UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL

The population genomics of western Iberian *Squalius* freshwater fish species: a genotyping by sequencing approach

Sofia Lopes Mendes

Mestrado em Biologia Evolutiva e do Desenvolvimento

Dissertação orientada por:
Professor Doutor Vítor Sousa
Professora Doutora Maria Manuela Coelho

2018

Acknowledgments

First, I would like to thank both my supervisors, for accepting me as their student, trusting me with this project and being the most amazing supervisors anyone can wish for. To professor Manuela, thank you for introducing me to the beauty of evolution and especially to the beauty of freshwater fish. To Vítor, thank you for the incredible amount of patience and for introducing me to the world of bioinformatics. Thank you both for your guidance, support, encouragement and for giving me the opportunity to learn so much and grow, both in science and as a person.

To all my colleagues in the Evolutionary Genetics group for providing such a great work environment and especially to João, Carlos and Cátia, who were conducting their master dissertations at the same time, for making this such an enjoyable year. I would also like to thank Miguel Machado for the helpful suggestions and discussion regarding the analysis of paired-end GBS data.

To all my friends, for their support and encouragement. In particular, to my friend Mariana, for sharing her accomplishments and frustrations with me and listening to mine, no matter how small some problems might seem now. Thank you for sharing this journey with me since the first day at FCUL. To my friend Margarida, for being such an example of positivity and perseverance. To my friend Marta, for always being here, throughout the good and the bad.

To Nuno, for being such a support. Thank you for your patience and encouragement and for believing in me when I didn't.

A toda a minha família, por ter sempre acreditado em mim. Ao meu pai, por trabalhar todos os dias para que eu pudesse chegar aqui. À minha mãe, por me ensinar sempre que podia chegar onde quisesse. À minha avó, sem a qual muito disto não seria possível. E, finalmente, à melhor irmã do mundo, que acredita sempre em mim e que me alegra e motiva todos os dias a ser melhor.

Abstract

In freshwater fish, processes of speciation and divergence between populations are, in many cases, extremely interconnected with the geomorphology of the rivers and lakes and the formation of geological barriers that can isolate populations. One geographical area where the isolation and the configuration of the drainage systems has been postulated to explain the high number of endemic fish species and their distributions is the Iberian Peninsula. Here, we focused on four species of the genus Squalius found in Portuguese rivers: *S. carolitertii*, *S. pyrenaicus*, *S. aradensis* and *S. torgalensis*. Previous genetic studies of these species using few mitochondrial (mtDNA) and nuclear markers revealed two main lineages: one comprising *S. aradensis* and *S. torgalensis* and another comprising *S. carolitertii* and *S. pyrenaicus*. Within the second lineage, incongruences were uncovered between mtDNA and nuclear markers. While on mtDNA phylogenies *S. pyrenaicus* formed a monophyletic group, for the nuclear markers populations from the Tagus river basin clustered with *S. carolitertii* instead of with other *S. pyrenaicus* populations (e.g. Guadiana basin). However, due to the limited number of markers, the processes underlying these incongruences could not be uncovered. Here, for the first time, we successfully obtained genome-wide single nucleotide polymorphisms (SNPs) for these species, using a Genotyping by Sequencing (GBS) approach without a reference genome. With this SNP dataset, we characterized the genetic diversity and differentiation patterns within and between species and inferred a species tree. Moreover, we investigated the possibility of introgression between *S. carolitertii* and *S. pyrenaicus* in the Tagus basin and modelled their demographic history. Our results uncovered two main lineages, in agreement with previous studies: one comprising *S. aradensis* and *S. torgalensis*, another comprising *S. carolitertii* and *S. pyrenaicus* from the Tagus basin as a sister clade to *S. pyrenaicus* from the Guadiana basin. Furthermore, this genome-wide dataset allowed us to detect and quantify introgression in the Tagus basin. Our estimates suggest that this lineage received a contribution from both *S. carolitertii* and *S. pyrenaicus* from Guadiana, although in different proportions.

Key-words: Iberian freshwater fish; *Squalius*; introgression; speciation; demographic modelling

Resumo

Compreender a divergência das populações e a formação de novas espécies é um dos objetivos principais da biologia evolutiva. Em peixes de água doce, estes processos estão muitas vezes relacionados com a formação das bacias hidrográficas, devido ao seu impacto no fluxo genético entre populações. Não só a dispersão destes animais está limitada aos rios e lagos que habitam, como alterações no curso e limite dos mesmos ao longo do tempo podem, por exemplo, isolar populações ou colocar em contacto espécies que evoluíram em alopatria. Uma região onde o isolamento de populações devido aos processos de formação das bacias hidrográficas parece ter contribuído para uma grande diversidade de espécies endémicas é a Península Ibérica. Um dos vários géneros presentes, caracterizado pela presença de espécies endémicas, é o género *Squalius*.

Em Portugal, excluindo um complexo híbrido *(Squalius alburnoides)*, estão descritas quatro espécies deste género: *S. carolitertii*, *S. pyrenaicus*, *S. torgalensis* e *S. aradensis*. Estas espécies têm distribuições discretas, habitando diferentes bacias hidrográficas. Esta elevada diversidade de espécies associada à distribuição em bacias distintas sugere que a sua evolução resultou de um processo de especiação alopátrica. Estudos genéticos anteriores revelaram a existência de duas linhagens: *S. aradensis* e *S. torgalensis*, que partilham um ancestral comum mais recente entre si do que com a segunda linhagem, onde se incluem *S. carolitertii* e *S. pyrenaicus*. No entanto, estes estudos revelaram também incongruências entre filogenias obtidas com marcadores nucleares e mitocondriais. Filogenias obtidas com marcadores mitocondriais indicam que *S. pyrenaicus* é monofilético, enquanto que as filogenias de genes nucleares indicam que *S. pyrenaicus* da bacia do Tejo é geneticamente mais próximos de *S. carolitertii* do que de *S. pyrenaicus* de outras bacias (Guadiana, Sado).

Neste trabalho, foram obtidos dados genómicos (SNPs) para estas quatro espécies através de um protocolo de "Genotyping by Sequencing (GBS)". Foi desenvolvida uma abordagem bioinformática de modo a permitir obter SNPs sem recurso a um genoma de referência. Para cada espécie, pelo menos uma bacia hidrográfica representativa dentro da sua distribuição foi amostrada. Com estes dados, foram caracterizados os padrões de diversidade e diferenciação genética e foi inferida uma filogenia das espécies e populações ("species tree"). Mais ainda, foram aplicados métodos para detetar e quantificar a introgressão entre *S. carolitertii* e *S. pyrenaicus* na bacia do Tejo.

Os resultados obtidos confirmam que os dados genómicos suportam a presença de duas linhagens, de acordo com estudos prévios: uma primeira linhagem composta por *S. aradensis* e *S. torgalensis* e uma segunda composta por *S. caroliterii* e *S. pyrenaicus*. Em todos os métodos aplicados, os indivíduos de *S. aradensis* e *S. torgalensis* agruparam claramente de acordo com a respetiva espécie. Mais ainda, *S. aradensis* e *S. torgalensis* apresentaram menor diferenciação genética (F_{ST}) entre si do que em relação às outras duas espécies. A aplicação de métodos de agrupamento com base nos padrões genómicos de cada indivíduo, como análise de componentes principais e inferência de proporções de ancestralidade, separam não só as duas linhagens (a de *S. carolitertii* e *S. pyrenaicus* e a de *S. torgalensis* e *S. aradensis*), como *S. aradensis* de *S. torgalensis* em dois grupos distintos das outras espécies.

Para a segunda linhagem, composta por *S. carolitertii* e *S. pyrenaicus*, tanto os métodos baseados em indivíduo como a diferenciação genética medida com F_{ST} calculada entre diferentes locais de amostragem mostram a existência de dois grupos: (i) um grupo formado por *S. carolitertii* e *S. pyrenaicus* da bacia do Tejo e de outra pequena bacia próxima (Lizandro) e (ii) outro grupo formado por *S. pyrenaicus* do Guadiana e pequenas bacias próximas (Almargem e Quarteira). Não só foram detetados maiores níveis de diferenciação genética entre *S. pyrenaicus* do Tejo e *S. pyrenaicus* do Guadiana do que entre *S. pyrenaicus* do Tejo e *S. carolitertii*, como os métodos baseados no indivíduo (análise de

componentes principais e inferência de proporções de ancestralidade) separaram claramente estes dois grupos. A filogenia das espécies e populações ("species tree") inferida indica que é possível separar as linhagens de *S. carolitertii* e *S. pyrenaicus* do Tejo, que, de acordo com a topologia inferida, partilham um ancestral comum após terem divergido da linhagem de *S. pyrenaicus* do Guadiana.

Para investigar se a maior proximidade genómica entre *S. carolitertii* e *S. pyrenaicus* do Tejo pudesse ser devida a introgressão, foram feitos testes utilizando a estatística D (também conhecida como teste ABBA/BABA). Quando foi considerada uma topologia em que *S. pyrenaicus* do Guadiana e do Tejo são taxa irmãos e *S. carolitertii* é a possível fonte de introgressão, foram obtidos valores de D significativamente positivos, indicando que *S. carolitertii* partilha mais alelos com *S. pyrenaicus* do Tejo do que com *S. pyrenaicus* do Guadiana. No entanto, outra alternativa é que *S. carolitertii* e *S. pyrenaicus* do Tejo têm um ancestral comum mais recente, e que os padrões são devidos a polimorfismo ancestral. A alteração da topologia colocando *S. carolitertii* e *S. pyrenaicus* do Tejo como taxa irmãos, com *S. pyrenaicus* do Guadiana como a fonte de introgressão, o valor de D obtido não foi significativamente diferente de zero, indicando que *S. carolitertii* e *S. pyrenaicus* do Tejo partilham aproximadamente o mesmo número de alelos com *S. pyrenaicus* do Guadiana, de acordo com a "species tree" inferida. Foi também investigada a possibilidade de locais de amostragem na bacia do Tejo mais próximos da distribuição de *S. carolitertii* mostrarem evidências de maior introgressão. No entanto, os resultados obtidos mostraram que este não é o caso, indicando que qualquer possível introgressão com *S. carolitertii* foi anterior à separação de *S. pyrenaicus* do Tejo em diferentes afluentes.

Considerando os resultados da estatística D, a explicação mais simples para a "species tree" inferida parece ser a de *S. carolitertii* e *S. pyrenaicus* do Tejo partilham um ancestral comum mais recente, após a divergência da linhagem de *S. pyrenaicus* do Guadiana, sem necessidade de invocar introgressão. No entanto, a realização de modelação demográfica baseada no espectro de frequências alélicas ("site frequency spectrum"), que utiliza mais informação do que a estatística D, indica que um cenário com fluxo genético é uma melhor explicação para os dados. De acordo com as estimativas obtidas, as populações de *S. pyrenaicus* do Tejo possuem uma contribuição no seu genoma de *S. caroliterti* (≈86%) e *S. pyrenaicus* do Guadiana (≈14%). Dois possíveis cenários podem explicar estes resultados. Uma hipótese é que *S. pyrenaicus* do Tejo tem na sua origem um evento de hibridação entre *S. carolitertii* e *S. pyrenaicus* do Guadiana. Neste caso, no passado, teria de existir contacto entre as paleobacias mais a norte, que deram origem às bacias onde *S. carolitertii* pode ser encontrado, e as paleobacias que deram origem ao Tejo e ao Guadiana, de modo a permitir fluxo genético entre *S. carolitertii* e *S. pyrenaicus* do Guadiana. A segunda hipótese é a de um contacto secundário, isto é, a linhagem de *S. pyrenaicus* do Tejo evoluiu independentemente, mas, devido a contactos que se possam ter restabelecido posteriormente entre bacias, ocorreu fluxo genético posterior à separação desta linhagem.

Os modelos simples testados neste trabalho não permitem distinguir entre as hipóteses de origem híbrida ou contacto secundário. No entanto, revelam a contribuição das linhagens de *S. carolitertii* e *S. pyrenaicus* do Guadiana no genoma de *S. pyrenaicus* do Tejo. Estes resultados sugerem, portanto, que a história evolutiva e especiação nestas espécies é provavelmente mais complexa do que um cenário de divergência sem fluxo genético associado à formação das bacias hidrográficas. Isto evidencia a necessidade de que mais modelos sejam testados, com uma maior amostragem que inclua bacias hidrográficas que não foram estudadas neste trabalho, de modo a poder distinguir entre estas duas hipóteses.

Palavras-chave: peixes de água doce da Península Ibérica; *Squalius*; introgressão; especiação; modelação demográfica

Table of Contents

List of Tables and Figures

Supplementary Figure S7 – Percentage of variance explained by each principal component (PC) on the Principal Components Analysis performed with a reduced dataset of one SNP per block (Supplementary Figure 6).

Supplementary Figure S8 - p-values of principal components on the PCA ($p<0.01$) performed with a reduced dataset of one SNP per block (Supplementary Figure 6).

Supplementary Figure S9 – Cross-entropy for each number of ancestral populations K when sNMF was performed on the dataset with only one SNP per block.

Supplementary Figure S10 – Ancestry proportions inferred with sNMF for four ancestral populations (K=4) for the dataset with one SNP per block.

Supplementary Figure S11 – Species tree graph obtained with TreeMix for the dataset with one SNP per block.

Supplementary Table S1 – Detailed sampling locations with GPS coordinates and fishing licences.

Supplementary Table S2 – Individual median depth of coverage after mapping all reads against the catalogue.

Supplementary Table S3 – Number of SNPs per sampling location that significantly deviate from Hardy-Weinberg equilibrium ($p<0.05$) due to a deficit (A) or excess (B) of heterozygotes for the different filtering options.

Supplementary Table S4 – Percentage of missing data of each individual on the final dataset.

Supplementary Table S5 - Number of polymorphic and monomorphic sites, missing data, private sites and fixed differences within each sampling locations.

Supplementary Table S6 – Mean expected heterozygosity and mean observed heterozygosity across sites for each sampling location.

Supplementary Table S7 – Quantiles 5% and 95% for the distribution of the expected and observed heterozygosity across sites for each sampling location.

Supplementary Table S8 - Number of polymorphic and monomorphic sites, missing data, private sites and fixed differences within each group.

Supplementary Table S9 – Mean expected heterozygosity and mean observed heterozygosity across sites for each group identified.

Supplementary Table S10 – Quantiles 5% and 95% for the distribution of the expected heterozygosity and observed heterozygosity across sites for each group identified.

Supplementary Table S11 – Detailed results of the D-statistic calculated for the different scenarios in Fig. 2.1.

Supplementary Table S12 - Detailed results of the D-statistic calculated per individual for the different scenarios in Fig. 2.1.

1. Introduction

Answering questions regarding how populations diverge and ultimately originate new species is a major goal of evolutionary biology. Speciation is assumed to occur due to a systematically reduction in gene flow through time until reproductive isolation is achieved and populations maintain phenotypic and genetic distinctiveness (Seehausen et al. 2014). The most acceptable hypothesis is that divergence happens in a strictly allopatric scenario, with absence of gene flow, due to barriers (geological, hydrological, etc.), which can lead to reproductive isolation due to the accumulation of genetic incompatibilities (Sousa and Hey 2013). However, there are now several studies based on phenotypic and genomic data suggesting that past gene flow is common in many species, even between populations/species adapted to different environments (Seehausen et al. 2014). It is also important to distinguish cases of divergence with gene flow during the divergence process from cases where populations first get isolated and then come into contact after a period of time and exchange migrants – a secondary contact scenario (Sousa and Hey 2013). Thus, to understand the process of speciation it is important to characterize the timing and mode of gene flow.

The ability to generate a massive number of polymorphic genetic markers scattered across the genome, due to the development of next-generation sequencing (NGS) technologies, has helped to shed light on the process of speciation and population divergence (Andrews et al. 2016). Reduced representations sequencing methods, like genotyping by sequencing (GBS), take advantage of NGS technologies to obtain single nucleotide polymorphisms (SNPs) spread across the genome (Davey et al. 2011). Population genomics approaches rely on the study of numerous loci, either generated by reduced representation or whole-genome sequencing, to gain insight into the processes that shape the evolutionary history of populations, attempting to quantify the influence of genetic drift, migration and different forms of natural selection (Luikart et al. 2003). This type of data can be used, for example, to quantify and characterize the genetic variability within and among populations and make inferences about population demographic history (Seehausen et al. 2014; Wolf and Ellegren 2016). This approach can be applied even when little previous genetic information is available, enabling the study of wild populations of non-model organisms (Narum et al. 2013). In this manner, valuable information to characterize the genetic diversity and differentiation of populations is generated, allowing to answer questions not only in evolutionary biology, but also in ecology and conservation biology (Narum et al. 2013). These methods have been employed in the study of speciation and the relationship between species in a multitude of organisms (e.g. butterflies (Dasmahapatra et al. 2012), gorillas (McManus et al. 2015), sawflies (Bagley et al. 2017)), including freshwater fish (Hohenlohe et al. 2010; Meier, Sousa, et al. 2017).

The diversity of freshwater fish is remarkable, considering there are almost as many species of ray-finned fish in freshwater as there are in the oceans, despite the obvious difference in the size of these habitats (Seehausen and Wagner 2014). A variety of scenarios have been described to explain the differentiation of freshwater fish populations: transitions from marine to freshwater habitats (Jones et al. 2012; Terekhanova et al. 2014), adaptation to extreme environments (Pfenninger et al. 2015), differentiation along water depth clines (Barluenga et al. 2006; Gagnaire et al. 2013). Along with these, freshwater fish differentiation and speciation is assumed to be related with the geomorphology of the rivers and lakes such as the formation of geological barriers that can isolate populations (Seehausen and Wagner 2014). Since they are restricted to freshwater systems, their dispersal is limited by the connections between different rivers/lakes, which can lead to the isolation of populations. However, this does not mean that currently geographically separated populations have always been isolated, since the configuration of river and lake systems can change over time. In fact, several studies document both past

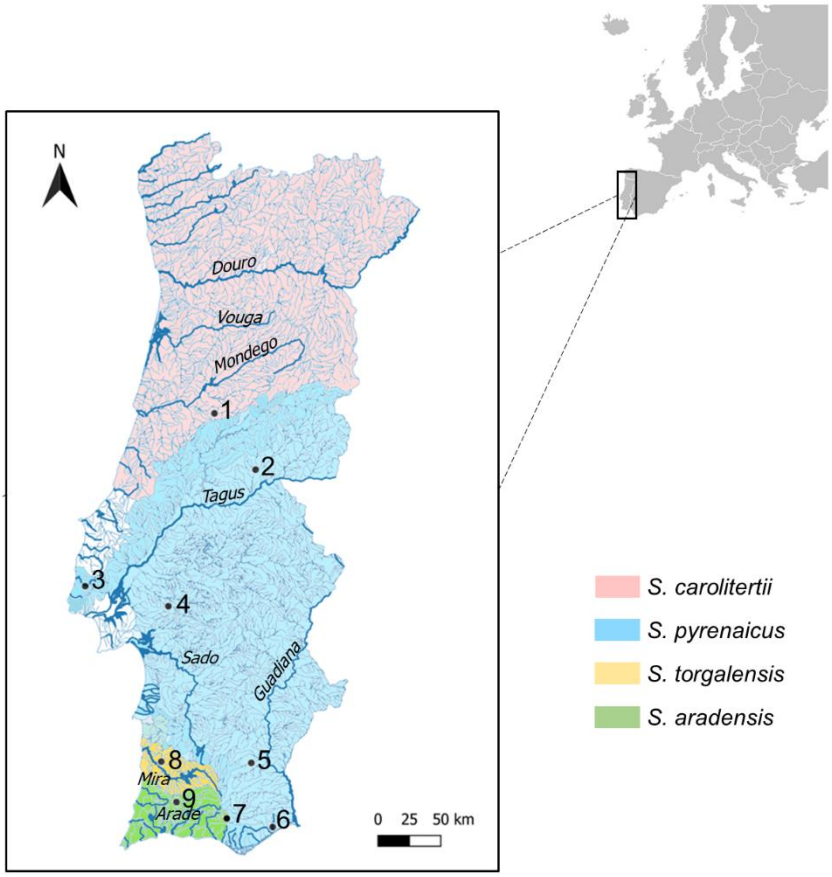


Figure 1.1 – Distribution range of the four *Squalius* species and sampling locations: (1) Mondego; (2) Ocreza; (3) Lizandro; (4) Canha; (5) Guadiana; (6) Almargem; (7) Quarteira; (8) Mira; (9) Arade.

and ongoing introgression in freshwater fish, both in species that have evolved with and without geographical isolation (Redenbach and Taylor 2002; Hohenlohe et al. 2013; Jones et al. 2013; Gante et al. 2016). Nonetheless, the role of reproductive isolation due to geographical barriers imposed by the geomorphology of the lakes and rivers remains a very important explanation for the abundance of freshwater fish species. One geographical area where isolation and the configuration of the drainage systems is assumed to be an important factor in the origin of a multitude of endemic fish species is the Iberian Peninsula.

The freshwater fish fauna of the Iberian Peninsula is highly diverse. Most of those belong to the family Cyprinidae, an extremely diverse family of freshwater fish, with representatives distributed through Eurasia, Africa and North America (Nelson et al. 2016). One of the several genera found in Iberian rivers is the genus *Squalius* Bonaparte, 1837 (sub-family Leuciscinae), commonly known as "chub". This is a key genus in Iberian ichthyofauna, with eight species and an hybrid complex currently described (Perea et al. 2016). These species belong to two different lineages: (1) the central European lineage, which includes only one of the species (*Squalius laietanus*), more closely related to species from central Europe than to the others on the peninsula, and (2) the Mediterranean lineage, to which all the other species belong to (Sanjur et al. 2003). All the *Squalius* species found in Portugal belong to this second lineage (Sanjur et al. 2003). Apart from the hybrid complex *Squalius alburnoides* (Steindachner, 1866), four species can be found in Portuguese rivers, all of them endemic to the Iberian Peninsula: *Squalius carolitertii*, *Squalius pyrenaicus*, *Squalius torgalensis* and *Squalius aradensis*. The distribution of these species is shown in Figure 1.1. *Squalius carolitertii* (Doadrio, 1988) is endemic to the northern region of the peninsula and its presence in Portugal was described from the northern border of the country to the Mondego basin (e. g. Coelho et al. 1995; Coelho et al. 1998). *Squalius pyrenaicus* (Gunther,

1868), on the other hand, has a more southern distribution range and is considered to be present in the Tagus, Sado and Guadiana basins (Coelho et al. 1995; Coelho et al. 1998). While these species have rather wide distribution ranges, the other two species found in Portugal are confined to much smaller river systems in the southwestern area of the country. *Squalius torgalensis* (Coelho, Bogustskaya, Rodrigues and Collares-Pereira, 1998) is endemic to the Mira river basin and *Squalius aradensis* (Coelho, Bogustskaya, Rodrigues and Collares-Pereira, 1998) can only be found in small drainages in the extreme southwestern area of the country (e.g. Arade, Seixe and Quarteira drainages) (Coelho et al. 1998). We note that in the Quarteira drainage, *S. pyrenaicus* is also present along with *S. aradensis* (Figure 1.1).

These species have been widely studied to understand their diversity, systematics and taxonomy (Brito et al. 1997; Sanjur et al. 2003; Mesquita et al. 2005; Henriques et al. 2010; Waap et al. 2011). Different approaches have already been employed, namely alloenzymes (Coelho et al. 1995), mitochondrial DNA (mtDNA) (Brito et al. 1997; Sousa-Santos et al. 2007) and nuclear markers such as microsatellites (Mesquita et al. 2005; Henriques et al. 2010) and nuclear genes (Waap et al. 2011). First and foremost, the initial studies using alloenzymes and mtDNA contributed to the establishment of the current taxonomy, supporting the recognition of the species level for the two wider distributed species (Coelho et al. 1995), and leading to the description of the two southwestern species (Coelho et al. 1998), in conjunction with the osteological data (Doadrio 1987; Coelho et al. 1998).

The relationship between these species described by mtDNA is consensual: phylogenies reconstructed with mitochondrial markers consistently indicated four well supported clades, corresponding to the four species (Brito et al. 1997; Sanjur et al. 2003; Mesquita et al. 2007), in accordance with the alloenzyme results (Coelho et al. 1995). These four species form a monophyletic group in the mtDNA phylogeny, even when other *Squalius* species from other locations around the Mediterranean are included in the phylogeny (Doadrio and Carmona 2003; Sanjur et al. 2003). Estimates based on fossil calibrations and markers from two mitochondrial and two nuclear genes date the most recent common ancestor of these four species to 14,6 Mya (Perea et al. 2010). Moreover, *S. torgalensis* and *S. aradensis* were found to be sister species, with the same being true for *S. carolitertii* and *S. pyrenaicus* (Brito et al. 1997; Sanjur et al. 2003). Molecular clock calibrations for cytochrome b (mtDNA) suggest an earlier differentiation of *S. torgalensis* and *S. aradensis*, compared to the divergence of *S. carolitertii* and *S. pyrenaicus*, occurring at a later stage (Brito et al. 1997; Sanjur et al. 2003). Overall, these mtDNA phylogenies coincided with the geographical distribution of the species in the different basins, therefore suggesting a pattern of allopatric speciation (Brito et al. 1997; Sanjur et al. 2003).

Interestingly, despite an apparent clear mtDNA phylogeny supported by independent studies, some incongruences were also found. For instance, in one study, the phylogeny reconstructed with cytochrome b (cyt b) showed that the majority of the individuals sampled from the Zêzere river, a tributary on the right margin of the Tagus, clustered with the northern species (*S. carolitertii*) instead of clustering with the other *S. pyrenaicus* from other rivers, including other Tagus tributaries (Brito et al. 1997). Furthermore, one *S. carolitertii* individual from the Mondego river clustered with *S. pyrenaicus* individuals instead of with the other *S. carolitertii* from Mondego (Brito et al. 1997). Although the authors pointed out the possibility of hybridization due to river captures and drainage direction changes of the Zêzere, and the hypothesis of possible fish translocations by man for fishing purposes in the Mondego, these incongruences were nevertheless considered to be exceptions, and the accepted interpretation was that species most likely evolved in isolation in the different drainages (Brito et al. 1997). Further studies based on mtDNA and beta-actin genes, while confirming the same relationship between the species and their distribution along the river basins, also reported the recovery of mitochondrial (cyt b) sequences from *S. carolitertii* in the Zêzere river (Sousa-Santos et al. 2007; Almada and Sousa-Santos 2010).

Regarding the genetic diversity of the species, mitochondrial diversity revealed a pattern of increasing genetic diversity from north to south, with *S. carolitertii* having the lowest mitochondrial variability (Brito et al. 1997; Sanjur et al. 2003), in accordance with previous alloenzyme data, where this lower genetic variability was also uncovered (Coelho et al. 1995). *S. pyrenaicus* seems to harbour the most genetic diversity at the mitochondrial level (Brito et al. 1997; Almada and Sousa-Santos 2010). The genetic diversity of *S. aradensis* appears to be variable between different areas of its distribution. The populations of this species are highly fragmented, and the levels of genetic diversity vary between the small rivers where it can be found (Mesquita et al. 2005). Considerable levels of genetic differentiation were reported between most populations, both at mitochondrial and nuclear (microsatellites) markers, possible as a consequence of habitat fragmentation (Mesquita et al. 2005). On the other hand, *S. torgalensis* showed very incipient population structure and lower genetic diversity than the mean of the *S. aradensis* populations for the same markers (Henriques et al. 2010).

The aforementioned studies were important and provided valuable information, not only for the understanding of the relationship between species, but also for conservation, since *S. pyrenaicus* is endangered and *S. aradensis* and *S. torgalensis* are classified as critically endangered. However, these studies never tried to infer a multi-locus species tree, obtaining and comparing only phylogenetic trees based on single genes or investigating patters of intraspecific variability and differentiation for conservation purposes. However, investigating the history of species based on single genes can be problematic, due to the highly stochastic effects of genetic drift and mutational processes (Hey and Machado 2003). For example, the information one can retrieve from mtDNA is limited as it actually behaves as a single locus due to the fact it is maternally inherited and lacks recombination (Allendorf 2017). When species diverged relatively recently, genes trees might not reflect the underlaying species tree (phylogeny) due to incomplete lineage sorting and/or gene flow. Given a species tree, the extent of variation in gene trees is described by coalescent theory, and the probability of incongruence with the species tree is affected by the split times, effective sizes and migration between populations (Hey and Machado 2003). Therefore, the topology of one particular gene tree might not be the same as the one of the species tree (Nichols 2001).

The first attempt to obtain a species tree for these four *Squalius* species was based on seven nuclear genes and produced some conflicting results with the previously consensual mitochondrial gene trees (Waap et al. 2011). This uncovered the same two main evolutionary lineages previously identified with mtDNA (e.g. Brito et al. 1997): the one of the southwestern species, *S. aradensis* and *S. torgalensis*, and another comprising *S. carolitertii* and *S. pyrenaicus* (Waap et al. 2011). However, in the nuclear DNA species tree (based on concatenating seven genes), *S. pyrenaicus* appears to be paraphyletic: *S. pyrenaicus* from Tagus and a small adjacent basin clustered with *S. carolitertii*, forming a sister clade to the remaining *S. pyrenaicus* from southern basins (Guadiana, Sado and Almargem) (Waap et al. 2011). This type of discordance between the nuclear and mitochondrial trees has been reported for other Iberian *Squalius* species outside of the Portuguese river basins (Perea et al. 2010). To explain the paraphyly of *S. pyrenaicus*, the authors point out the possibility that the separation from *S. carolitertii* in Mondego and *S. pyrenaicus* in the Tagus could be a recent event, possibly related to the development of the current drainage system in the later Pliocene/Pleistocene (Waap et al. 2011). Moreover, the configuration of the southern basins (Sado and Guadiana), which were connected during the Miocene, might explain why *S. pyrenaicus* from the south are more closely related to each other than to Tagus (Waap et al. 2011). Nonetheless, in Waap et al. 2011 only individuals from one tributary on right margin of the Tagus were sampled, so it was impossible to determine if only individuals from tributaries on the right margin were genetically closer to *S. carolitertii* or if the pattern was consistent throughout the Tagus basin. Moreover, although seven nuclear genes constitute an improvement over phylogenies strictly based on mitochondrial DNA, this still provides a limited picture of the genome.

Therefore, this work had three major goals: (i) first, to characterize the genome-wide patterns of genetic diversity and differentiation; (ii) second, to reconstruct the species tree for these four *Squalius* species in Portuguese river basins; and (iii) to investigate the possibility of introgression between *S. carolitertii* and *S. pyrenaicus* in order to solve the incongruences left by previous work. To achieve these goals, we successfully obtained genome-wide single nucleotide polymorphisms (SNPs) through a Genotyping by Sequencing (GBS) protocol and developed a pipeline to analyze GBS data from different species without a reference genome.

2. Methods

Sampling and sequencing

A total of 96 individuals were sampled from nine different locations, as displayed on Figure 1.1. For each species, at least one sampling location from a representative drainage system was sampled. For *S. carolitertii*, individuals were collected from the Mondego basin (n=10). For *S. pyrenaicus*, in the northern part of its distribution individuals were collected from the Ocreza river (n=10) and Ribeira de Canha (n=10), both tributaries of the Tagus basin. Specimens were also collected in the Lizandro basin (n=10). From here on, "northern *S. pyrenaicus*" refers to *S. pyrenaicus* from Ocreza, Canha and Lizandro as a whole. In the southern part of the distribution, *S. pyrenaicus* was sampled in the Guadiana (n=2), Almargem (n=12) and Quarteira basins (n=10). From here on, the designation "southern *S. pyrenaicus*" refers *S. pyrenaicus* from Guadiana, Almargem and Quarteira as a whole. For *S. aradensis*, individuals were collected from the Arade (n=5) and Quarteira basins (n=17). Finally, *S. torgalensis* individuals were collected in the Mira basin (n=10). Detailed locations with GPS coordinates and fishing licenses from Instituto de Conservação da Natureza e das Florestas (ICNF) can be found on Supplementary Table S1.

All fish were collected by electrofishing (300V, 4A), and total genomic DNA was extracted from fin clips using an adapted phenol-chloroform protocol (Miller et al. 1988). DNA was quantified using Qubit® 2.0 Fluorometer (Live Technologies). The samples were subjected to a paired-end Genotyping by Sequencing (GBS) protocol (adapted from Elshire et al. 2011), performed in outsourcing at Beijing Genomics Institute (BGI, www.bgi.com). The DNA samples were sent to the facility mixed with DNAstable Plus (Biomatrica) to preserve DNA at room temperature during shipment. Briefly, upon arrival, DNA was fragmented using the restriction enzyme ApeKI and the fragments were amplified after adaptor ligation (Elshire et al. 2011). The resulting library was subjected to Illumina Hiseq2000 sequencing. All these steps had already been performed by other members of the Evolutionary Genetics group upon my arrival, and hence the raw sequence data was already available for analysis.

Obtention of a high-quality SNP dataset

The obtention of SNPs and individual genotypes from the raw data was a large component of this work and is detailed in this section. To aid in the comprehension of the pipeline, a schematic representation is displayed in Figure 2.1. First, the quality of the sequences of each individual was assessed using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). To compile the information from all individual reports, we used MultiQC (Ewels et al. 2016) to merge and summarize the individual FastQC reports. Second, we used the program *process_radtags* from Stacks version 2.0 (Catchen et al. 2013) to trim all reads to 82 base pairs and discard reads with low quality scores. A sliding window of 0.15x the length of the read was used to eliminate reads where the average quality score drops below a defined threshold of 10 (in phred score). The default settings were used for the size of the window and the quality threshold. We verified the success of the trimming and cleaning by obtaining new quality reports for the clean reads using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) followed by MultiQC (Ewels et al. 2016) as before. Given the absence of a reference genome for any of the species in study, in order to map clean reads and identify single nucleotide polymorphisms (SNPs), we built a reference catalogue of all loci using a *denovo* assembly approach. To do so, we used Stacks version 2.0 (Catchen et al. 2013), which is tailored to deal with short-reads generated by various reduced-representation sequencing protocols.

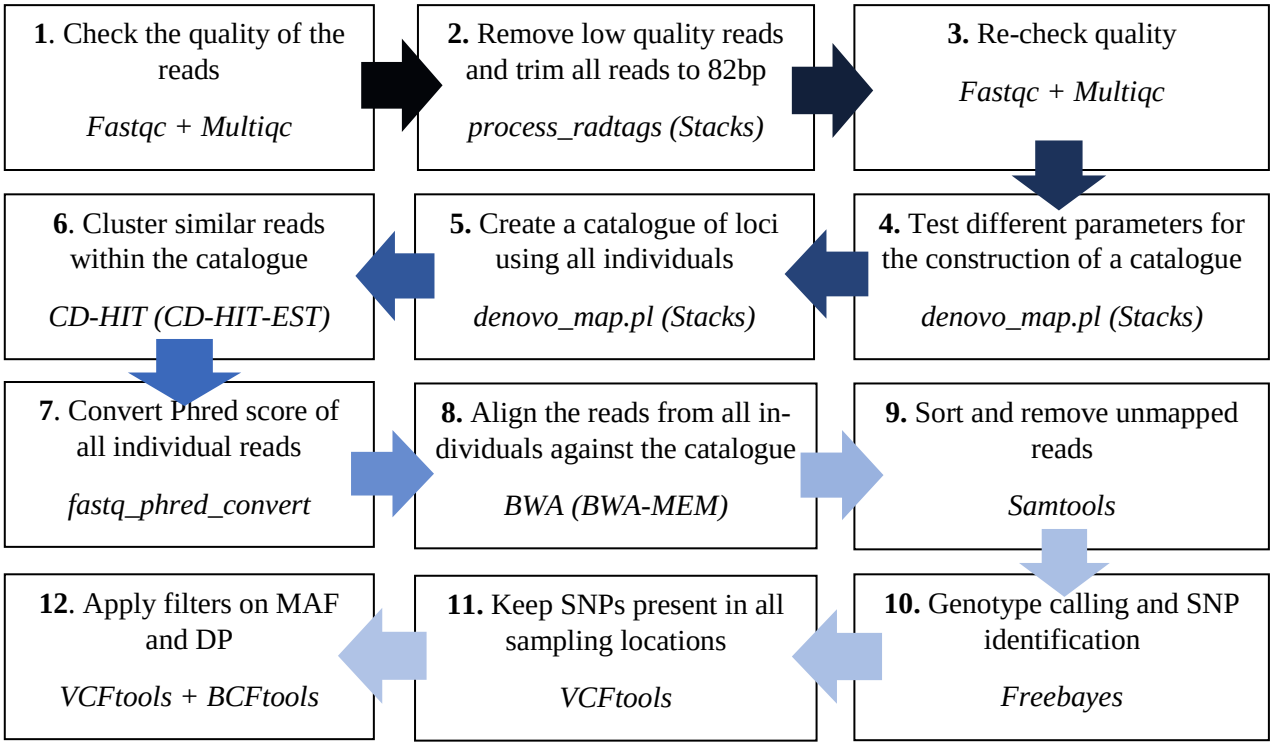


Figure 1.1 – **Schematic representation of the pipeline followed to obtain a dataset of SNPs based on paired-end GBS data from different species.**

For the denovo assembly, the method implemented in Stacks identifies loci present across different individuals and builds a catalogue (reference) of all loci. The construction of the catalogue depends on a set of key parameters: (i) *m*, the minimum coverage of each loci; (ii) *M*, the maximum number of differences between sequences within the same individual for them to be considered from the same locus; (iii) *n*, the maximum number of differences between sequences from different individuals for them to be considered from the same locus on the catalogue. There is thus a trade-off when deciding the value of the *M* and *n* parameters: allowing many mismatches can lead to two sequences from different genomic regions to be wrongly mapped to same locus, which can increase the number of SNPs; on the other hand, allowing for few differences can lead to wrongly considering alleles from the same locus as different loci, which can diminish the number of SNPs. The choice of these parameters should take into consideration the diversity level of the species in study. To determine the best values for the parameters, we followed the approach recommended by Paris et al. 2017. We used a small subset of 15 individuals representative of all species to test different values for the parameters. We tested values of *M* between 1 and 8, while keeping n = 1. For *n*, values between 1 and 10 were tested while keeping *M*=2. In both cases, *m* was kept as default (*m*=3) and all SNPs were required to be present in all species (*p*=5) in at least 50% of the individuals (*r*=0.5). Considering the trade-off mentioned above and the results we obtained (see below), we decided to use *M*=4 and *n*=4, while keeping *m* as default (*m*=3), for the construction of the catalogue.

We had to adapt the *denovo* method implemented on Stacks for our paired-end GBS data. Stacks was designed primarily for RADseq protocols (ex. Baird et al. 2008) and the differences in library preparation from these type of protocols to a paired-end GBS protocol create problems in the analysis (see Davey et al. 2011 for a comparison of the protocols). In RADseq, the adaptors of two paired-ends are different, which is not the case for GBS, where the two ends of the fragment are not distinguishable. For this reason, the algorithm implemented in Stacks could wrongly treat as different loci the forward and reverse sequences. To merge the forward and reverse reads and eliminate duplicate loci from the catalogue, similar reads within the catalogue were clustered using CD-HIT version 4.7 (Li and Godzik

2006; Fu et al. 2012). CD-HIT-EST from the CD-HIT package was used with a word length of 8 and a sequence identity threshold of 0.85. Shortly, this method clusters sequences based on identity by sorting sequences from long to short, considering the longest sequence as the first representative and going through each sequence classifying it as a new representative or redundant, in comparison with the list of representative sequences already found.

Once we obtained a clean catalogue, this was treated as a reference genome and the reads from each individual were aligned against it using BWA-MEM from BWA version 0.7.17-r1188 (Li 2013) with default parameters. Before this step, the quality scores of the individual reads had to be converted, since sequencing was performed using Illumina 1.5 encoding, and therefore base quality in Phred quality scores were encoded using ASCII + 64. We converted the scoring to ASCII + 33 before the alignment with BWA using the program *fastq_phred_convert*, available online (https://github.com/greatfireball/fastq_phred_convert). This was done because ASCII + 64 was discontinued and hence recent tools expect ASCII + 33 (e.g. BWA). Performing the conversion at this step of the pipeline does not pose a problem because the programs used before either allow to specify the base quality score (e.g. Stacks) or do not require quality scores (e.g. the catalogue used on CD-HIT). The output alignments of BWA were then sorted and unmapped reads were removed using Samtools version 1.8 (Li et al. 2009). Then, to call genotypes for each individual at each site and identify SNPs, we used the method implemented on Freebayes v1.2.0 (Garrison and Marth 2012). Provided with a reference (in this case, the catalogue) and mapped reads for different individuals, Freebayes detects genetic variants and infers genotypes for each individual at each variant site outputting them in VCF format. VCFtools version 0.1.15 (Danecek et al. 2011) was then used to keep only SNPs, as Freebayes also detects other small polymorphisms (e.g. indels). All SNPs were required to be present in all sampling sites in at least 50% of the individuals, using a combination of options from VCFtools (Danecek et al. 2011).

To discard sites and genotypes that are more likely to be the result of sequencing or mapping errors, we applied filters on the minor allele frequency (MAF $\geq$ 0.01) and on depth of coverage (DP). The aim of these filters was to remove from the dataset sites with rare variants (MAF filter) and genotypes with low or high DP, such that only genotypes with a DP near the expected median were kept. Given that the depth of coverage varied considerably among individuals, rather than applying the same filter to all individuals, two filtering options were explored: (A) treat as missing data genotypes with a DP outside of ⅓ to 2x the individual median DP; (B) treat as missing data the genotypes with a DP outside of ¼ to 4x the individual median DP. For both (A) and (B), we performed a Hardy-Weinberg test to remove SNPs with an excess of heterozygotes ($p<0.01$) when pooling all individuals in the dataset, as high heterozygosity in a given SNP can be re result of mapping errors (e.g. duplicated regions mapped the same locus). The different filters were applied using a combination of options from VCFtools version 0.1.15 (Danecek et al. 2011) and BCFtools version 1.6 (Li et al. 2009). We assessed the effect of these filtering options on the number of SNPs and percentage of missing data. We also calculated the number of SNPs that significantly deviate from Hardy-Weinberg equilibrium within each sampling location ($p<0.05$), either by a deficit or excess of heterozygotes, after each filtering option. The filters that provided the best compromise between the number of SNPs obtained and the percentage of missing data were chosen, and that dataset was used for all further analysis. Finally, we calculated the percentage of missing data per individual for the chosen dataset.

Characterization of the global patterns of genetic diversity and differentiation

Once a high-quality SNP dataset was obtained, the global patterns of genetic diversity and differentiation within and among species were evaluated. First, we calculated, within each sampling

location, the number of (i) polymorphic sites; (ii) monomorphic sites; (iii) sites with data for only one individual; (iv) private sites, i.e. sites that are polymorphic in one population but monomorphic on the others; (v) sites with fixed differences, i.e. all individuals in one population are homozygotes for one allele in that site but all individuals from the other populations are homozygotes for the other allele. To evaluate the levels of genetic diversity in each sampling location, the mean expected and observed heterozygosity were calculated. Moreover, to quantify the levels of differentiation between locations, we calculated the pairwise F_{ST} using the Hudson estimator (Hudson et al. 1992). Given that the sampling locations may not correspond to populations, we investigated fine population structure with individual-based methods. To understand how individuals cluster, we conducted a principal component analysis (PCA). The number of significant principal components was determined with the Tracy-Widom test (Patterson et al. 2006) on all eigenvalues. Furthermore, individual ancestry proportions were estimated with the sparse Non-negative Matrix Factorization method (sNMF) (Frichot et al. 2014). This method infers the best number (K) of ancestral populations to explain the data, as well as the proportion of each individual's genome assigned to each of K "populations" (ancestry proportions) (Frichot et al. 2014). We tested values of K between 1 and 10, performing 100 repetitions for each K value. We then calculated the mean expected and observed heterozygosity for each of the identified clusters, as well as the pairwise F_{ST} between them, in the same manner as before. All calculations were performed in RStudio version 1.1.383 and R version 3.4.4 using custom scripts, and the PCA and sNMF were done using the package LEA (Frichot and François 2015).

Inference of a population and species tree

Given that our sampling included different species and populations within species, we used the SNP data to reconstruct a species and population tree describing the relationships between the populations using TreeMix (Pickrell and Pritchard 2012). This program implements a maximum likelihood method and models the changes in allele frequencies due to genetic drift with a Gaussian approximation. Assuming that all sites are independent, it allows to infer the topology of the relationships among populations as the graph that best fits the variance and covariance of allelic frequencies among populations (Pickrell and Pritchard 2012). We explored a scenario with no migration, as well as models allowing for up to two migration events. Since we do not have an outgroup, the position of the root was not specified, and thus the resulting trees are unrooted.

Effect of linked SNPs

It is noteworthy that PCA, sNMF and TreeMix methods assume that SNPs are independent, and thus results can be affected by linked SNPs. Here, given the absence of a reference genome, we lack information on the location of the SNP markers. It is thus difficult to evaluate linkage disequilibrium (LD) patterns and detect linked SNPs, except for sites within the same scaffold of the catalogue. To verify if the results were influenced by potential linkage of SNP markers, we produced a dataset by dividing the catalogue into blocks of 200 base pairs and selecting only one SNP per block. For each block we selected the SNP with the less missing data. Using this single SNP dataset, we repeated the PCA, sNMF (Frichot et al. 2014) and TreeMix (Pickrell and Pritchard 2012) analysis.

Detection of introgression between *S. carolitertii* and *S. pyrenaicus*

To test for possible past introgression between *S. carolitertii* and *S. pyrenaicus* in the northern area of *S. pyrenaicus* distribution, we used the D-statistic (Durand et al. 2011), also known as ABBA/BABA test. This test distinguishes ancestral polymorphism from gene flow by looking at

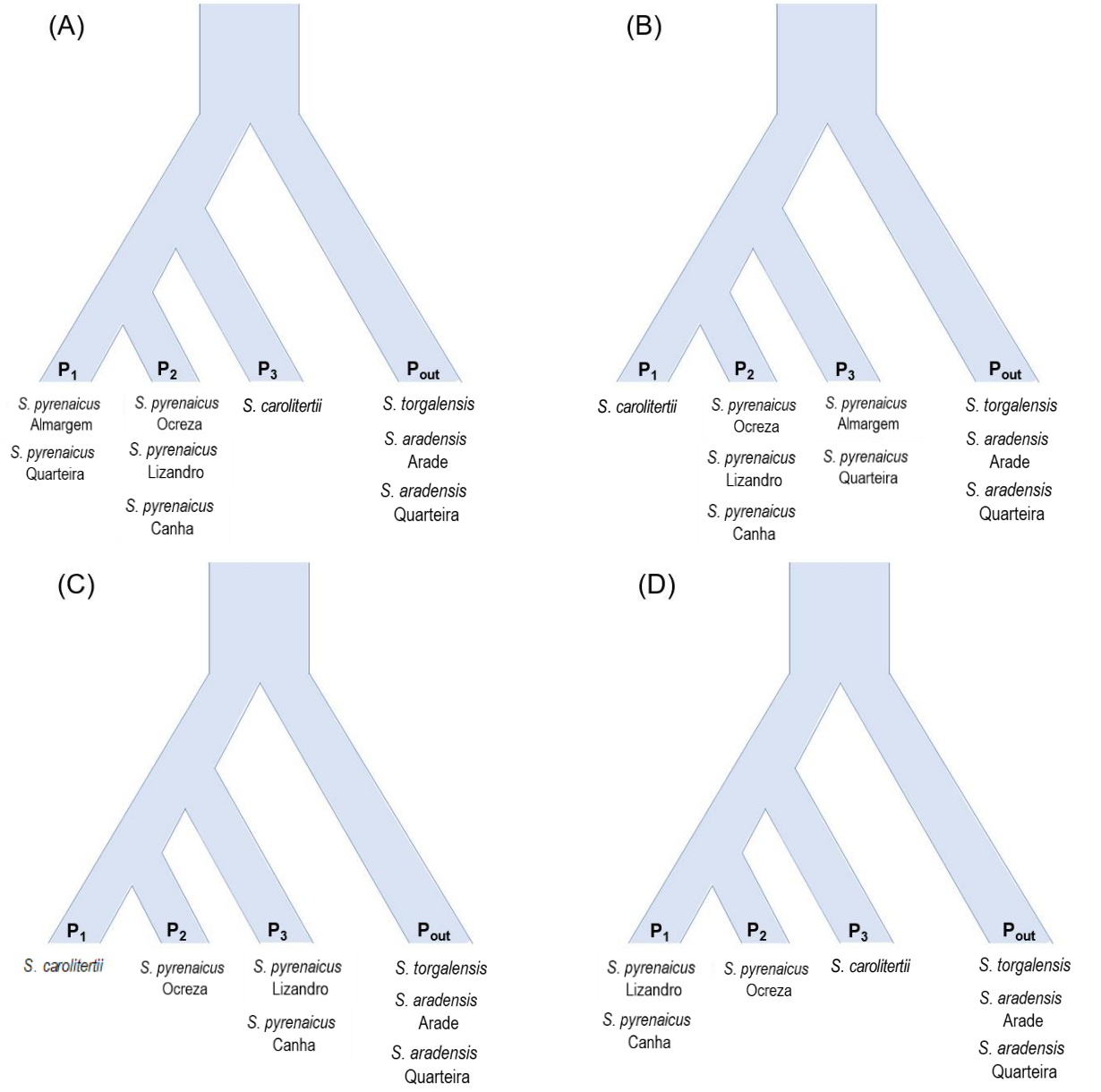


Figure 2.2 – Different species trees explored with D-statistic. For each species tree considered, the value of D was calculated for all possible combinations of populations, which are indicated below each population P1, P2, P3 and Pout.

incongruences between the gene trees and the species/population tree. To implement this test, it is required to have data from four different populations related through a fixed species tree: two sister populations (P1 and P2), a third population that could be the source of introgressed genes (P3) and has a common ancestor to P1 and P2, and one outgroup (Pout). If we define the ancestral state in the outgroup as A and the derived state in the third population as B and focus on the SNPs where the two sister populations have different alleles, there are only two possibilities: ABBA or BABA, where the order of the alleles refers to the population order (P1, P2, P3, Pout). These correspond to sites where there is an incongruence between the gene trees and the population/species tree. If ancestral polymorphism is the cause of the incongruences (i.e. no introgression), we expect P3 to be equally distant from P1 and P2 and thus the number of SNPs with the ABBA and BABA pattern to be identical. This is expected because if the two alleles were already present in the ancestral of the three populations (P1, P2 and P3) then both P1 and P2 are equally likely to share the derived allele with P3. In this case of ancestral polymorphism (incomplete lineage sorting) the value of the D-statistic will be zero. On the other hand, if there is introgression (gene flow) between one of the sister populations (P1 and P2) and the third population

(P3), we expect an excess of SNPs with the ABBA or BABA pattern and D will be significantly different from zero. In that case, if there is an excess of SNPs with the ABBA pattern and a significant positive D-statistic, it indicates gene flow between P2 and P3. Otherwise, if there is an excess of SNPs with the BABA patterns and a significant negative D-statistic, it indicates gene flow between P1 and P3.

Here, we explored four different possible species trees to perform different tests, as shown in Figure 2.2. In A, we tested for introgression between *S. carolitertii* (P3) and two sister populations (P1 and P2) from *S. pyrenaicus*, one from the northern and another from the southern part of its distribution. In B, we tested if *S. pyrenaicus* populations from the south (P3) are more closely related to *S. carolitertii* (P1) or populations from the northern part of *S. pyrenaicus* distribution (P2). Considering the possibility of a geographical cline in admixture proportions between *S. carolitertii* and *S. pyrenaicus* in the northern part of *S. pyrenaicus* distribution, we also tested if the northern most sampling site of *S. pyrenaicus* (Ocreza – see Figure 1.1) showed more signs of introgression with *S. carolitertii* than the other northern *S. pyrenaicus*, which corresponds to scenario C. The opposite (all northern *S. pyrenaicus* as sister populations and *S. carolitertii* as the potential source of introgressed genes) corresponds to scenario D. In all cases, the outgroup (Pout) was either *S. torgalensis* or *S. aradensis*. All possible combinations of the populations shown in the figure were tested. *S. pyrenaicus* Guadiana was deliberately left out as there are only two individuals form this sampling location and one of them has a very high percentage of missing data (see results). Significance of D-statistic values was assessed using a jackknife approach, dividing the dataset into 25 blocks and converting z-scores into p-values assuming a standard normal distribution ($p<0.01$).

If introgression between populations occurred in the relatively recent past, we would expect individuals within the same population to show different degrees of introgression. To test this hypothesis, we calculated the D-statistic per individual of P2 for the same scenarios (Fig. 2.2). All possible combinations were tested, and significance of D was assessed using the jacknife approach as described above.

Demographic modelling of the divergence of *S. carolitertii* and *S. pyrenaicus*

We compared alternative divergence scenarios of the northern *S. pyrenaicus* from *S. carolitertii* and the southern *S. pyrenaicus* to test and quantify past introgression events. We used the composite likelihood method based on the joint site frequency spectrum (SFS) implemented in *fastsimcoal2* (Excoffier et al. 2013). We compared the fit of two models to the observed SFS: no admixture and admixture (Figure 2.3). The admixture model assumes that the northern *S. pyrenaicus* received a contribution alpha (α) from the southern *S. pyrenaicus* and 1-alpha (1-α) from *S. carolitertii* at the time of the split. Note that the estimates of alpha not only indicate the most likely species tree but also quantify the level of introgression. If alpha=0 then the northern *S. pyrenaicus* is more closely related to *S. carolitertii*, whereas if alpha=1 then the northern and southern *S. pyrenaicus* are closer to each other. Values of alpha in between 0 and 1 indicate that the northern *S. pyrenaicus* received a contribution from both species, and hence indicate introgression. To test if a model with admixture fits better the observed data, we compared the likelihood of this admixture model to a model without admixture, i.e. with alpha=0. To be able to compare the likelihood values directly, models need to have the same number of parameters. Thus, in the model without admixture we allowed for a bottleneck associated with the split of the northern *S. pyrenaicus* from *S. carolitertii*, mimicking a founder effect. All parameters were scaled in relation to a reference effective size, which was arbitrarily set to be the *Ne* of *S. carolitertii*. To obtain an observed SFS without missing data, we built the joint 3D-SFS by sampling 4 individuals from *S.*

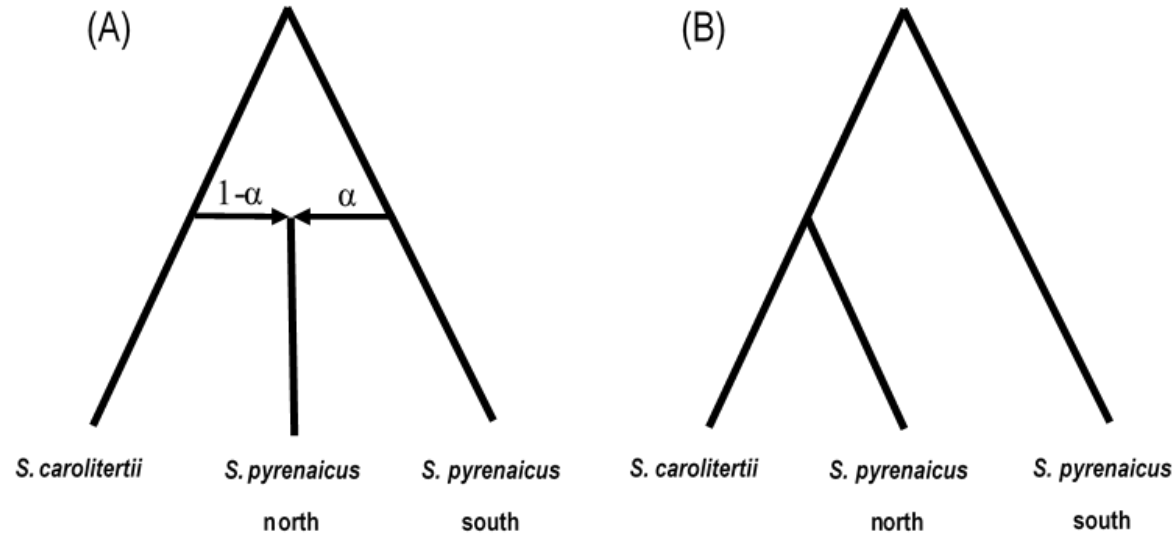


Figure 2.3 – Schematic representation of the two models compared with *fastsimcoal2*. (A) admixture; (B) no admixture. To compare directly the log10(likelihood), both models have the same number of parameters. The admixture model assumes that at the time of divergence the northern *S. pyrenaicus* received a contribution α from the southern *S. pyrenaicus* and a contribution 1-α from *S. carolitertii*. The model with no admixture has α=0, but to have the same number of parameters we allowed for a bottleneck associated with the divergence of *S. pyrenaicus*, mimicking a founder effect.

carolitertii and the southern *S. pyrenaicus*, and 6 individuals from the northern *S. pyrenaicus*. Given the lack of an outgroup, we could not identify the ancestral state of alleles, and hence used the minor allele frequency spectrum. To sample individuals without missing data, we used the initial dataset but without the MAF filter, and each scaffold was divided into blocks of 200bp (which is larger than the average length of the GBS loci), and for each block we sampled the individuals from each population with less missing data keeping only the sites with data across all individuals. Given that the SFS is affected by the depth of coverage, only genotypes with DP>10 were used (Nielsen et al. 2011). This resulted in an observed SFS with 8900 SNPs. For each model we performed 50 independent runs with 50 cycles, approximating the SFS with 100,000 coalescent simulations.

3. Results

Obtention of a high-quality SNP dataset

After the initial processing of the reads, removing low quality reads and trimming all reads to 82 base pairs, we obtained a mean of 5,223,433 high quality reads per individual. These reads were used to construct the *denovo* assembly catalogue. Regarding the selection of the parameters for the construction of the catalogue, we found an overall increase in the number of SNPs when allowing for higher number of mismatches, both within the same individual (*M* parameter) and between individuals (*n* parameter) (Supplementary Figures S1 and S2). When keeping the other parameters fixed, increasing the maximum number of differences between reads within the same individual (*M*) led to a slight increase in the number of SNPs for *M* between 2 and 8 (Supplementary Figure S1). When varying the maximum number of differences between reads from different individuals (*n*), the number of SNPs increased from *n*=1 to *n*=10, but the increment became smaller as *n* values increased (Supplementary Figure S2). We followed the recommendation of Paris et al. 2017 of keeping the value of *n* between *M*-1 and *M*+1 and chose a conservative value of *M*=4 and *n*=4 to create the final catalogue. This was done to maximize the number of SNPs, while minimizing the probability of wrongly treating different alleles from the same locus as different loci, and of wrongly treating similar or paralogous genomic regions as a single locus. After mapping all the reads from each individual to the catalogue, the median depth of coverage per sample was 37.5x. We note, however, that there was a large variation in the depth of coverage (DP) across individuals (Supplementary Table S2). In general, *S. aradensis* samples, as well as *S. pyrenaicus* from Quarteira, exhibited lower median DP than the majority of the individuals in the other samples.

Table 3.1 – Number of SNPs and percentage of missing data for the different filtering options.

	No filters	**MAF ≥ 0,01**	**MAF ≥ 0,01 + ⅓ to 2x median DP**	**MAF ≥ 0,01 + ¼ to 4x median DP**
Number of SNPs	42902	28257	27914	27703
% missing data	0	0	58.67%	42.39%

After obtaining a dataset of SNPs found in all sampling locations in at least 50% of the individuals in each location, we applied a filter on the minor allele frequency (MAF) and depth of coverage (DP). Table 3.1 shows the effect of the different filtering options on the number of SNPs and the resulting overall percentage of missing data (per individual and per site). Prior to the application of any filters, we obtained a total of 42,902 SNPs. Applying a filter on minor allele frequency, keeping only sites with MAF ≥ 0.01, decreased the number of SNPs to 28,257, suggesting that many SNPs contained rare alleles that can be due to sequencing errors. Further application of filters on DP had a much smaller effect on the number of SNPs, only decreasing by 343 (⅓ to 2x individual median DP) or 554 SNPs (¼ to 4x DP median DP) the size of the dataset. However, these DP filters had an effect on the percentage of missing data, producing datasets with ≈59% and ≈42% of missing data respectively.

For each sampling location, we assessed the number of SNPs that significantly deviate from Hardy-Weinberg equilibrium ($p<0.05$) with the different filtering options, both for deficit (Figure 3.1 A) and excess of heterozygotes (Figure 3.1 B). The precise values can be consulted on Supplementary Table S3. Overall, few SNPs deviate significantly from Hardy-Weinberg equilibrium in each sampling site. These deviations can be due to artefacts, such as sequencing errors (e.g allele dropout leading to

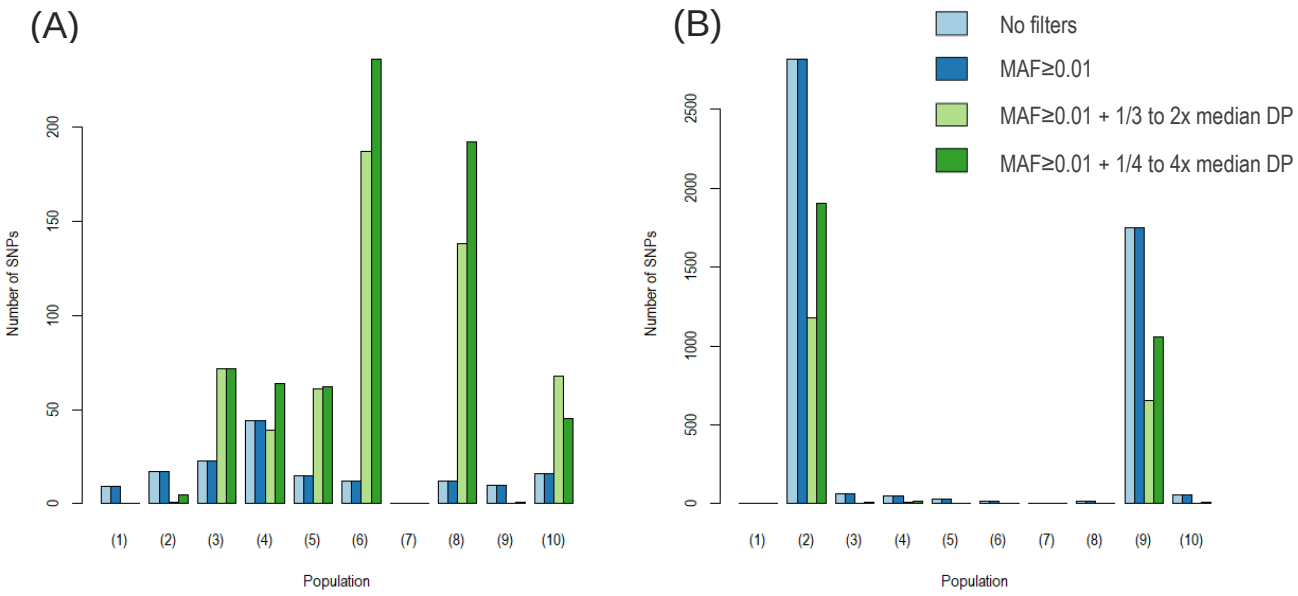


Figure 3.1 - Number of SNPs per sampling location that significantly deviate from Hardy-Weinberg equilibrium (p<0.05) due to a deficit (A) or excess (B) of heterozygotes for the different filtering options. Each sampling location is coded by a number: (1) *S. aradensis* Arade; (2) *S. aradensis* Quarteira; (3) *S. carolitertii*; (4) *S. pyrenaicus* Almargem; (5) *S. pyrenaicus* Canha; (6) *S. pyrenaicus* Lizandro; (7) *S. pyrenaicus* Guadiana; (8) *S. pyrenaicus* Ocreza; (9) *S. pyrenaicus* Quarteira; (10) *S. torgalensis*.

homozygote excess) and mapping errors (e.g. mapping duplications to the same location resulting in excess of heterozygotes), or biological factors, such as inbreeding, population structure or natural selection. Even using a conservative significance level of 0.05, without correcting for multiple tests, the number of SNPs that shows a significant deviation is particularly small when compared to the total number of SNPs on the dataset (Table 3.1), indicating that most sites are at Hardy-Weinberg equilibrium and that there is no evidence for genome-wide effects of inbreeding and population structure within each sampling site. Appling filters based on MAF does not remove any SNPs significantly out of Hardy-Weinberg equilibrium (Figure 3.1), although it decreased the size of the dataset (Table 3.1). As expected if regions with very high depth of coverage were in part due to mapping errors, resulting in wrongly calls of heterozygote genotypes, the filters applied on DP decreased the number of SNPs with a significant excess of heterozygote individuals (Figure 3.1 B). It is noteworthy that the sampling location of Quarteira clearly stands out, as both species sampled in that location exhibit the highest number of SNPs with an excess of heterozygotes and that number is clearly much higher than in any other sampling location.

Considering the above results on the effects of different filters, for further analysis we decided to use the dataset filtered with MAF≥0.01 together with a filter on DP, keeping only genotypes with ¼ to 4x the individual median DP. We chose this option as it produced a dataset with lower missing data, while removing rare alleles that are likely sequencing errors (MAF filter) and discarding sites with very low quality (low DP) or that are result of mapping errors (very high DP), accounting for variation in depth of coverage per individual.

In the final dataset (filtered with MAF≥0.01 + ¼ to 4x the individual DP median), eighteen individuals have more than 50% of missing data (Supplementary Table S4). Of those, only six of them have a percentage of missing data higher than 60% but no individuals had more than 70% missing data. These individuals were distributed across sampling locations, rather than clustered in a single location. Given that individuals with a very high percentage of missing data were, at most, one per sampling site,

with the exception of Almargem where there where two, we decided to keep all individuals in the final dataset.

Characterization of the global patterns of genetic diversity and differentiation

Although sampling locations might not correspond to populations, we quantified the genetic diversity patterns at each location. The number of SNP sites across sampling location showed two different patterns (Supplementary Table S5). Both species sampled in Quarteira, as well as *S. aradensis* Arade and *S. pyrenaicu*s Guadiana, show the highest missing data (highest number of sites without data) but also the highest number of fixed differences. The remaining sampling locations have more SNPs (i.e. lower missing data) but less fixed differences and more monomorphic sites (Supplementary Table S5).

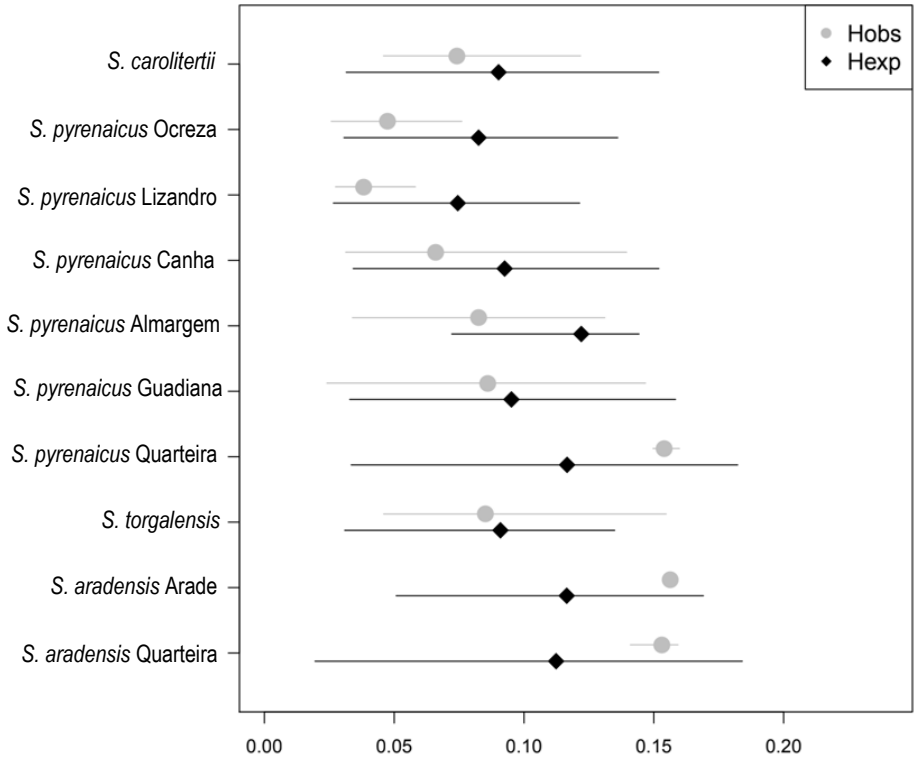


Figure 2.2 – Mean expected and observed heterozygosity for each sampling location. The lines represent the variation from quantile 5% to quantile 95% of the distribution across sites. Black diamonds represent the expected and grey circles the observed heterozygosity.

The mean observed and expected heterozygosity were similar across different sampling locations, with a large variation in the expected heterozygosity distribution across sites (Figure 3.2 and Supplementary Tables S6 and S7). Nevertheless, the two species sampled in Quarteira, as well as *S. aradensis* Arade show the highest levels of genetic diversity, with higher values of mean expected heterozygosity (Figure 3.2 and Supplementary Table S6). Within *S. pyrenaicus*, the south (Almargem, Guadiana, Quarteira) seems to harbour more genetic diversity than the north (Ocreza, Lizandro, Canha), as populations in the north have lower expected heterozygosity.

The pairwise F_{ST} estimates of genetic differentiation between sampling locations are shown in Table 3.2. The two southwestern species (*S. aradensis* and *S. torgalensis*) are less differentiated from each other than from *S. carolitertii* and *S. pyrenaicus*. Moreover, both south-western species are as differentiated from *S. carolitertii* as they are from *S. pyrenaicus* (Table 3.2). Interestingly, they are most

Table 3.2 – F_{ST} calculated between the different sampling locations. Colours correspond to those of the species distribution on Figure 1.1.

	S. carolitertii	*S. pyrenaicus* Ocreza	*S. pyrenaicus* Lizandro	*S. pyrenaicus* Canha	*S. pyrenaicus* Guadiana	*S. pyrenaicus* Almargem	*S. pyrenaicus* Quarteira	*S. torgalensis*	*S. aradensis* Arade	*S. aradensis* Quarteira
S. carolitertii	-	0.117	0.140	0.092	0.107	0.204	0.213	0.376	0.371	0.385
S. pyrenaicus Ocreza	-	-	0.124	0.075	0.126	0.219	0.236	0.392	0.387	0.400
S. pyrenaicus Lizandro	-	-	-	0.076	0.150	0.240	0.254	0.413	0.406	0.418
S. pyrenaicus Canha	-	-	-	-	0.091	0.194	0.201	0.372	0.367	0.380
S. pyrenaicus Guadiana	-	-	-	-	-	-0.090	-0.138	0.293	0.292	0.306
S. pyrenaicus Almargem	-	-	-	-	-	-	0.047	0.387	0.381	0.392
S. pyrenaicus Quarteira	-	-	-	-	-	-	-	0.371	0.360	0.367
S. torgalensis	-	-	-	-	-	-	-	-	0.221	0.249
S. aradensis Arade	-	-	-	-	-	-	-	-	-	-0.054
S. aradensis Quarteira	-	-	-	-	-	-	-	-	-	-

differentiated from *S. pyrenaicus* Lizandro (F_{ST}>0.4). For *S. aradensis*, there is no sign of genetic differentiation between the two sampling locations of this species (Arade and Quarteira). For that reason, they show similar levels of differentiation to all the other sampling locations, although the F_{ST} values for *S. aradensis* Quarteira are systematically higher (Table 3.2). Concerning *S. carolitertii* and *S. pyrenaicus*, these species are more differentiated from *S. aradensis* and *S. torgalensis* than from each other. However, it is evident that *S. carolitertii* is not equally differentiated from all *S. pyrenaicus* populations. Indeed, the levels of differentiation between S. carolitertii and northern *S. pyrenaicus* (Ocreza, Lizandro, Canha) are lower (F_{ST}<0.140) than between *S. carolitetii* and *S. pyrenaicus* from the south (Almargem, Quarteira) (F_{ST}>0.20) (Table 3.2). Interestingly, among the northern *S. pyrenaicus* sampling locations, *S. carolitertii* is more differentiated from Lizandro (Table 3.2). Within *S. pyrenaicus*, the northern sampling locations (Ocreza, Lizandro, Canha) appear to be more differentiated from the other *S. pyrenaicus* in the south (Almargem and Quarteira) (F_{ST}>0.194) than they are from *S. carolitertii* (F_{ST}<0.14) (Table 3.2). Among the southern *S. pyrenaicus* sampling locations, Guadiana clearly stands out as an outlier. Although it does not show any differentiation from the other southern *S. pyrenaicus* (Almargem and Quarteira), the F_{ST} values indicate is it more differentiated from the northern *S. pyrenaicus* from Lizandro and Ocreza than it is from *S. carolitertii* (Table 3.2). However, this result might be a consequence of the fact there are only two individuals sampled in Guadiana and one of them has a very high percentage of missing data (≈69.27%).

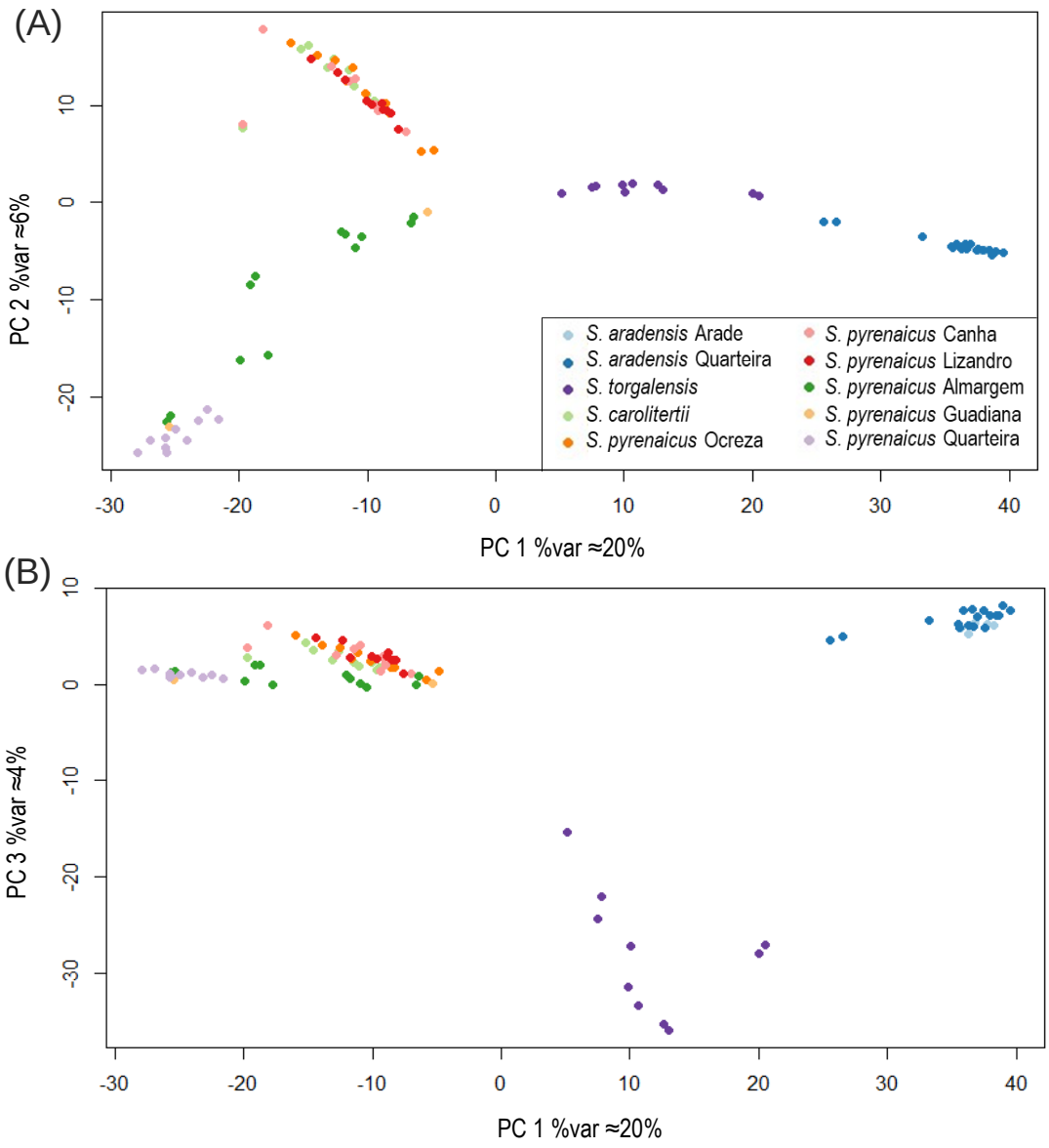


Figure 3.3 – Results for the first three components of the Principal Components Analysis: (A) PC1 and PC2; (B) PC1 and PC3; (C) PC2 and PC3. Each point corresponds to one individual. PC1 separates *S. aradensis* and *S. torgalensis* from *S. carolitertii* and *S. pyrenaicus*. PC2 separates the southern *S. pyrenaicus* from the northern *S. pyrenaicus* and *S. carolitertii*. Lastly, PC3 separates *S. aradensis* from *S. torgalensis*.

The PCA results show that the first three principal components explain approximately 30% of the variation (Supplementary Figure S3), although the Tracy-Widom tests (Patterson et al. 2006) indicate that the first six components have a significant effect ($p<0.01$) (Supplementary Figure S4). We only show the first three PCs because these have a clear biological interpretation. The first principal component (Figure 3.3 A – PC1) explains the higher percentage of the variance (≈20%) and clearly separates two groups: one formed by *S. carolitertii* and *S. pyrenaicus* (negative side of the axis) and another formed by *S. aradensis* and *S. torgalensis* (positive side of the axis). This is consistent with the pairwise F_{ST} results. The second principal component (PC2) explains a much lower percentage of the variance (≈6%) and mostly affects *S. carolitertii* and *S. pyrenaicus* separating them into two groups: one formed by *S. carolitertii* and northern *S. pyrenaicus* from Ocreza, Lizandro and Canha (all individuals above zero on the PC2 axis) and a second group formed by the southern *S. pyrenaicus* from Almargem, Guadiana and Quarteira (all individuals below zero on the PC2 axis). Plotting the first and third principal

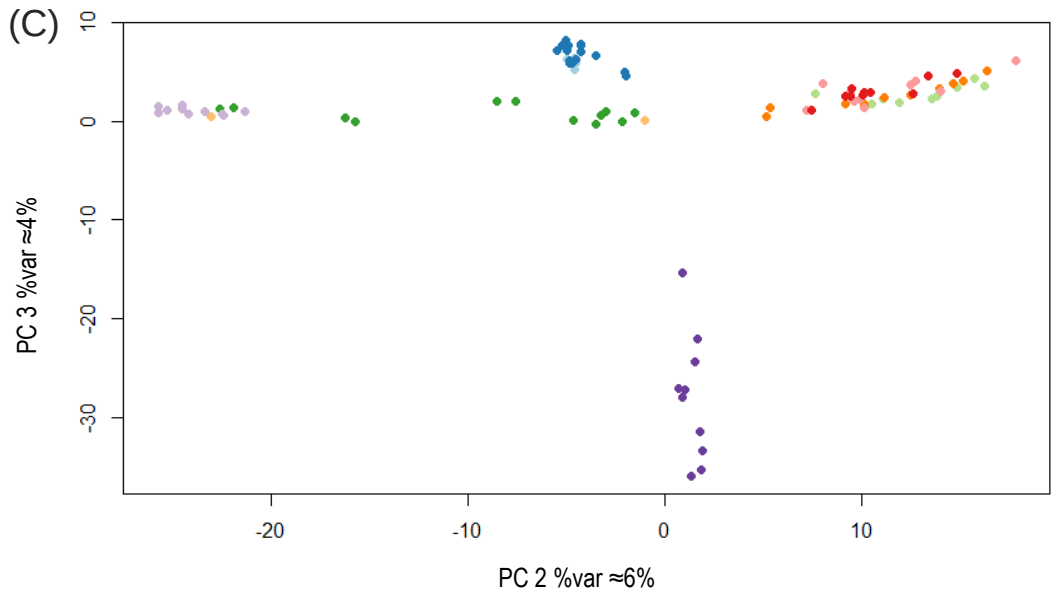


Figure 3.3 (cont.) – Results for the first three components of the Principal Components Analysis: (A) PC1 and PC2; (B) PC1 and PC3; (C) PC2 and PC3. Each point corresponds to one individual. PC1 separates *S. aradensis* and *S. torgalensis* from *S. carolitertii* and *S. pyrenaicus*. PC2 separates the southern *S. pyrenaicus* from the northern *S. pyrenaicus* and *S. carolitertii*- Lastly, PC3 separates *S. aradensis* from *S. torgalensis*.

components together, three clusters are evident (Figure 3.3 B). The separation between *S. aradensis* + *S. torgalensis* (positive side of PC1 axis) and *S. carolitertii* + *S. pyrenaicus* (negative side of PC1 axis) created by the first principal component is still visible but adding the third principal component separates *S. aradensis* (above zero on the PC3 axis) from *S. torgalensis* (below zero on the PC3 axis). Interestingly, when the second and third principal components are plotted together, the separation created by the second component on *S. carolitertii* and *S. pyrenaicus* is even more evident, creating a gradient along the PC2 axis (Figure 3.3 C). On the more negative end is *S. pyrenaicus* Quarteira, followed by one individual of *S. pyrenaicus* Guadiana and *S. pyrenaicus* Almargem more towards zero. We note that in the PCA, missing data is replaced by the mean allele frequency at each site, and thus individuals with higher missing data are driven to values close to zero (Patterson et al. 2006). This is indeed the case, as the individual with 69% missing data from *S. pyrenaicus* Guadiana is close to zero, which can also explain the pairwise F_{ST} results that indicated this sampling location exhibited lower levels of differentiation from *S. carolitertii* than the other southern *S. pyrenaicus* locations (Almargem and Quarteira). On the positive side of the PC2 axis there is a cluster formed by individuals from *S. carolitertii* and *S. pyrenaicus* from Ocreza, Lizandro and Canha, with no clear separation between individuals of different sampling locations (Figure 3.3 C).

The estimation of ancestry proportions and the mostly likely number of clusters with sNMF (Frichot et al. 2014) suggests that the data are consistent with four populations, with K=4 having the smallest cross-entropy value (≈0.2969) (Supplementary Figure S5). Although, K=5 appears to be an equally good fit for the data based on its cross-entropy value, we note that it is higher (≈0.2972) than K=4 and the result has no clear biological interpretation. The result for K=4 is displayed on Figure 3.4 and is consistent with the PCA and pairwise F_{ST} results. From left to right on the figure, the first cluster (light blue) contains the individuals from both sampling sites of *S. aradensis* (Arade and Quarteira). The second cluster (dark blue) includes the individuals from *S. torgalensis*. The third cluster (light green) comprises individuals from *S. carolitertii* and the northern sampling locations of *S. pyrenaicus* (Canha, Lizandro and Ocreza). The final cluster includes individuals from *S. pyrenaicus* Quarteira, *S. pyrenaicus* Guadiana and *S. pyrenaicus* Almargem. Different individuals from *S. pyrenaicus* Almargem seem to

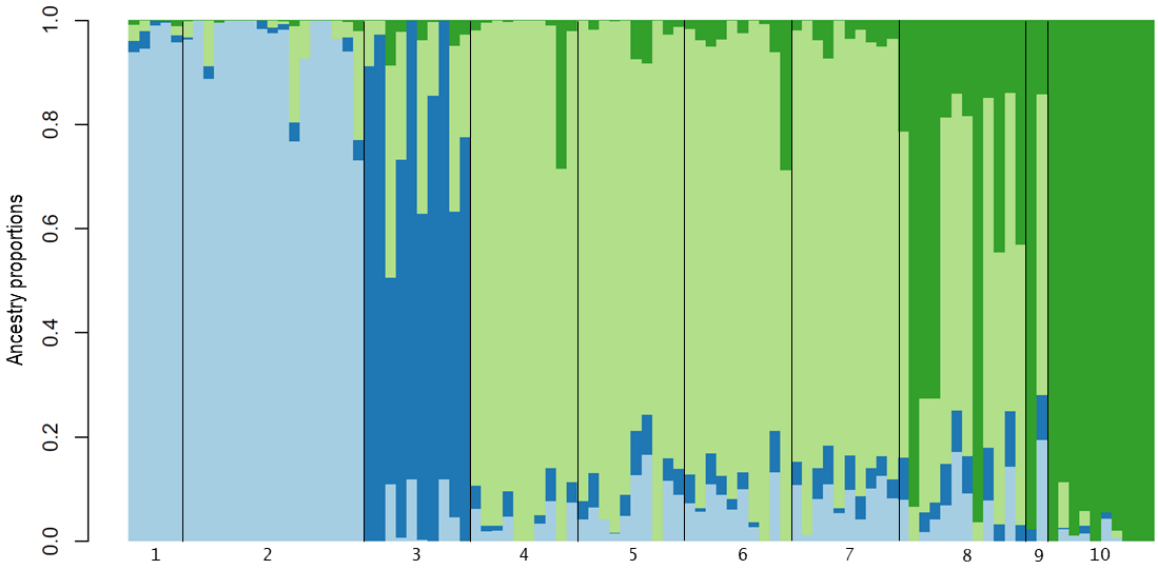


Figure 3.4 – Ancestry proportions inferred with sNMF for four ancestral populations (K=4). Each vertical bar corresponds to one individual and the proportion of each colour corresponds to the estimated ancestry proportion from a given cluster. The individuals are grouped per sampling locations and the groups are separated by black lines. Each number corresponds to a sampling location: (1) *S. aradensis* Arade; (2) *S. aradensis* Quarteira; (3) *S. torgalensis*; (4) *S. carolitertii*; (5) *S. pyrenaicus* Ocreza; (6) *S. pyrenaicus* Canha; (7) *S. pyrenaicus* Lizandro; (8) *S. pyrenaicus* Almargem; (9) *S. pyrenaicus* Guadiana; (10) *S. pyrenaicus* Quarteira.

cluster either within the third or fourth cluster (some have a high proportion of light green while others are almost totally dark green). This is also the case for *S. pyrenaicus* Guadiana, where one individual is totally assigned to cluster four but not the other. However, methods like sNMF are sensitive to missing data and that might influence these results. Moreover, virtually all individuals in the dataset exhibit some small proportion from groups other than the one they are assigned to, which can be due to statistical noise or shared ancestral polymorphism.

Based on the PCA and sNMF results, we pooled individuals into four groups: (i) *S. aradensis*, (ii) *S. torgalensis*, (iii) *S. carolitertii* and the northern *S. pyrenaicus* and (iv) southern *S. pyrenaicus*. We found that the cluster formed by *S. aradensis* and the one of the southern *S. pyrenaicus* have the highest genetic diversity, reflecting the levels of diversity calculated per sampling location (Figure 3.5 and Supplementary Table S9). As for comparisons across sampling locations, we found high variance in the expected heterozygosity distribution across sites in the four groups (Figure 3.5 and Supplementary Table S10).

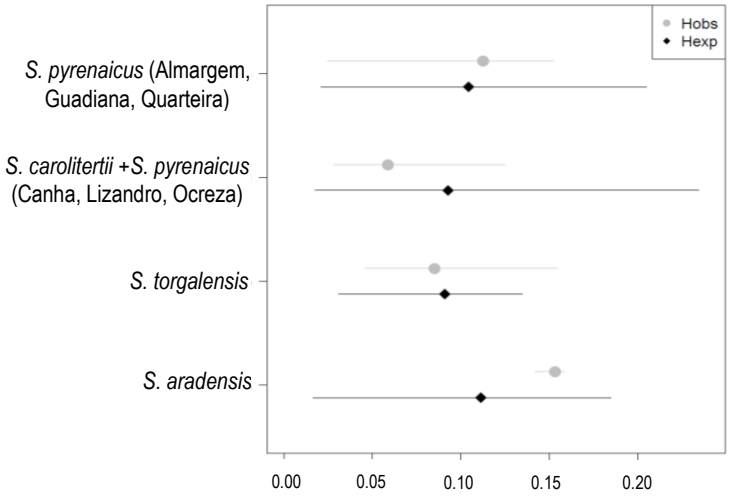


Figure 3.5 – Mean expected and observed heterozygosity for the four inferred clusters. The lines represent the variation from quantile 5% to quantile 95% of the distribution across sites. Black diamonds represent the expected and grey circles the observed heterozygosity. Clusters are in the same order as in Figure 3.4.

Table 3.3 – F_{ST} calculated between the four clusters identified with sNMF and PCA. Clusters are in the same order as in Fig. 3.3 (left to right).

	S. aradensis	***S. torgalensis***	***S. carolitertii* + *S. pyrenaicus* Canha + *S. pyrenaicus* Lizandro + *S. pyrenaicus* Ocreza**	***S. pyrenaicus* Almargem + *S. pyrenaicus* Guadiana + *S. pyrenaicus* Quarteira**
S. aradensis	-	0.252	0.374	0.380
S. torgalensis	-	-	0.360	0.375
***S. carolitertii* + *S. pyrenaicus* Canha + *S. pyrenaicus* Lizandro + *S. pyrenaicus* Ocreza**	-	-	-	0.180
***S. pyrenaicus* Almargem + *S. pyrenaicus* Guadiana + *S. pyrenaicus* Quarteira**	-	-	-	-

In agreement with the F_{ST} calculated per sampling location and the results of the PCA, we found that the groups of *S. aradensis* and *S. torgalensis* are genetically less differentiated from each other than they are from the other two groups (F_{ST}>0.361). The pairwise F_{ST} results also indicate that the group of *S. carolitertii* and northern *S. pyrenaicus* and the group of the southern *S. pyrenaicus* are less differentiated from each other than from the groups of *S. aradensis* or *S. torgalensis* (Table 3.3).

Inference of a population and species tree

We inferred a species tree based on the covariance of allele frequencies across all SNPs, modelling changes in allele frequencies through time due to genetic drift using TreeMix (Pickrell and Pritchard 2012). The inferred topology of the relationships between populations (unrooted tree) and the branch lengths (longer branches represent stronger genetic drift) is shown on Figure 3.6. This unrooted

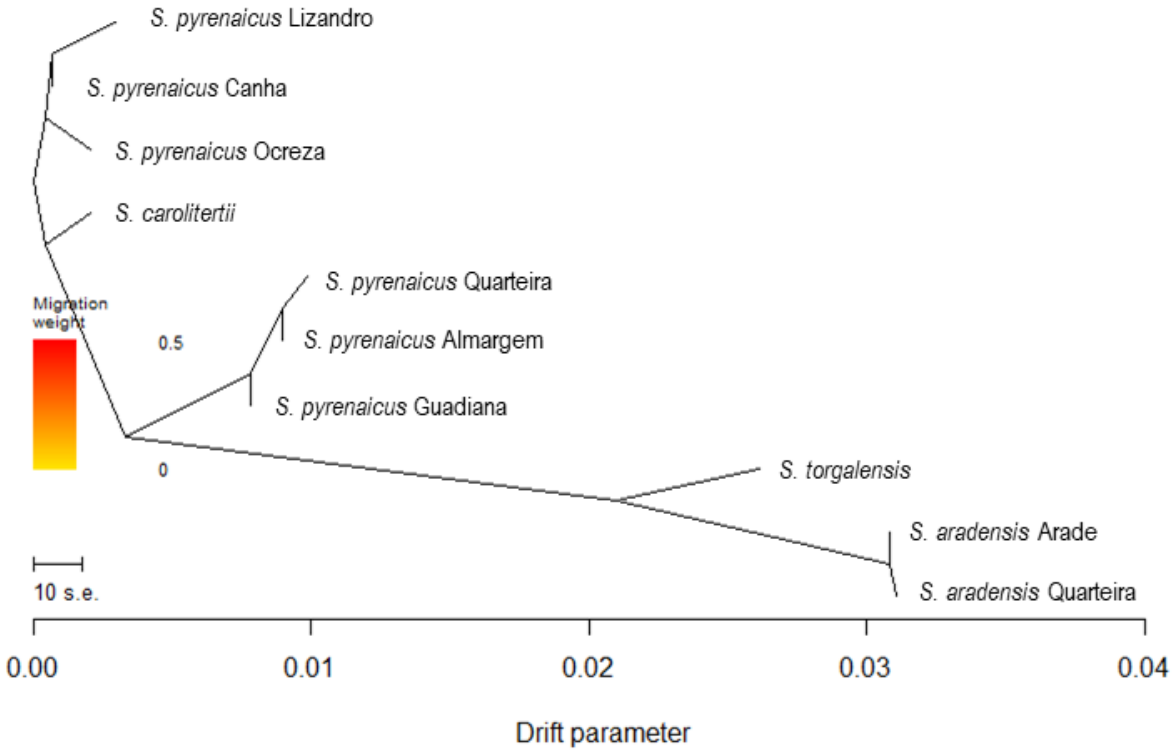


Figure 3.6 – Species tree graph obtained with TreeMix. This is an unrooted tree and branch lengths are represented in units of genetic drift, i.e. the longer a given branch the stronger the genetic drift experienced during that branch, which could be due to longer divergence times and/or smaller effective sizes.

tree shows a clear separation between two groups: one comprising *S. aradensis* and *S. torgalensis* and the other comprising *S. carolitertii* and *S. pyrenaicus*. *S. aradensis* and *S. torgalensis* appear as sister species, which is in accordance with the F_{ST}, PCA and sNMF results.

Within the lineage of *S. carolitertii* and *S. pyrenaicus*, two main groups are found: one formed by *S. pyrenaicus* from the south (Almargem, Guadiana and Quarteira) and another formed by *S. carolitertii* and northern *S. pyrenaicus* (Canha, Lizandro and Ocreza). This is in agreement with the PCA and sNMF, where these two clusters were also detected, as well as with the F_{ST} results that indicated a lower level of differentiation between northern *S. pyrenaicus* populations (Canha, Lizandro and Ocreza) and *S. carolitertii* than between northern and southern *S. pyrenaicus* populations. The longer branch lengths of the northern *S. pyrenaicus* (Canha, Lizandro and Ocreza) indicate that these populations went thought higher levels of genetic drift when compared to the southern *S. pyrenaicus* (Almargem, Guadiana and Quarteira). This is also in accordance with the lower expected heterozygosity calculated for these populations.

Attempts to produce a species tree with one or two migration events were unsuccessful as different runs of TreeMix did not produce consistent results (different migration events on different runs) and thus could not be trusted. Therefore, only results without migration are shown, as these were consistent for different runs of the program.

Effect of linked SNPs

To verify if the results were influenced by the fact that some SNPs could be linked, we produced a dataset with only one SNP per block of 200 base pairs. This dataset comprised 4,220 SNPs and the overall percentage of missing data was ≈43.69 %. When repeating the PCA, sNMF and the TreeMix analysis, all results were consistent with those from the initial dataset of 27,703 SNPs. Regarding the PCA, the percentage of the variance explained by the first three components slightly increased to ≈35% (Supplementary Figure S7). Although the Tracy-Widom tests (Patterson et al. 2006) indicate that the first seven components have a significant effect ($p<0.01$) (Supplementary Figure S8), we only show the first three PCs because these have a clear biological interpretation. As for the initial dataset, the first PC separated the *S. aradensis* and *S. torgalensis* from *S. carolitertii* and *S. pyrenaicus*, the second PC further differentiated *S. carolitertii* and the northern *S. pyrenaicus* from the southern *S. pyrenaicus* and the third PC separated *S. aradensis* from *S. torgalensis* (Supplementary Figure S6). Concerning the results of sNMF (Frichot et al. 2014), we found again that K=4 was also the best number of clusters to describe the data (Suplementary Figure S9) and the same individuals were placed within each cluster – one cluster comprising *S. aradensis*, a second cluster comprising *S. torgalensis*, a third cluster comprising *S. carolitertii* and the northern *S. pyrenaicus* (Ocreza, Lizandro and Canha) and a fourth cluster comprising the southern *S. pyrenaicus* (Supplementary Figure 10). Finally, the topology of the species tree obtained with TreeMix, was also identical to the one previously obtained (Suplementary Figure S11), with the same two main groups (*S. torgalensis* + *S. aradensis* and *S. carolitertii* + *S. pyrenaicus*). As before, a closer proximity between the northern *S. pyrenaicus* and *S. carolitertii* than between the northern and southern *S. pyrenaicus* is evident. The consistency between these results and the ones previously obtained indicates that our results are not influenced by the possibility that some SNPs are linked. Hence further analysis were done using the initial dataset.

Detection of introgression between *S. carolitertii* and *S. pyrenaicus*

The results for the D-statistic (ABBA/BABA) calculated per population are displayed on Figure 3.7. The exact number of SNPs that showed the ABBA or BABA pattern and p-values can be found on

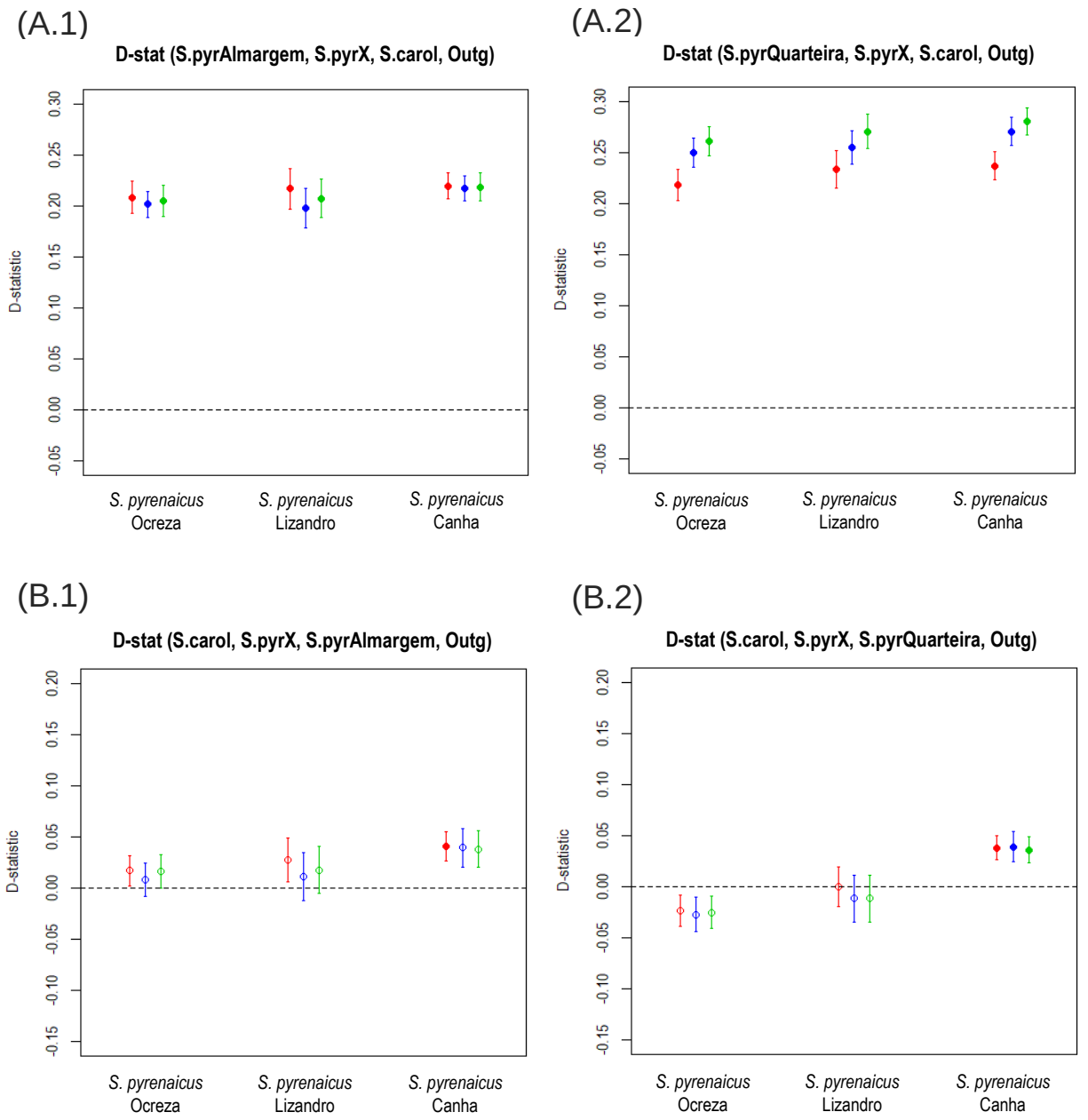


Figure 3.7 – Results of the D statistic calculated for the different scenarios in Fig. 2.1. For each topology (A to D), the results are presented according to the northern *S. pyrenaicus* population (S. pyrX) used. In A and B, two plots are presented, one for each southern *S. pyrenaicus* population used: Almargem (S.pyrAlmargem) and Quarteira (S.pyrQuarteira). "S.carol" stands for *S. carolitertii*, S.pyrOcreza stands for *S. pyrenaicus* Ocreza and "Outg" for outgroup. The result using each outgroup is coded with a different colour. Full dots represent significant D values ($p<0.01$).

Supplementary Table S11, where significant p-values are highlighted in grey shading. The results are shown in the same order as Figure 2.2 (Figure 3.7 A corresponds to the topology on Figure 2.2 A, and so on).

To test for introgression between *S. carolitertii* and *S. pyrenaicus*, we used the first topology (Figure 2.2 A). The resulting values of D were significantly positive for all population combinations (Figures 3.7 A.1 and A.2), independently of the southern *S. pyrenaicus* population (*S. pyrenaicus* Almargem (A.1) or *S. pyrenaicus* Quarteira (A.2)) used as P3 and of the outgroup used (*S. torgalensis* (red), *S. aradensis* Arade (blue) or *S. aradensis* Quarteira (green)). However, the value of D increased when *S. pyrenaicus* Quarteira was used as the representative of southern *S. pyrenaicus* P3 population,

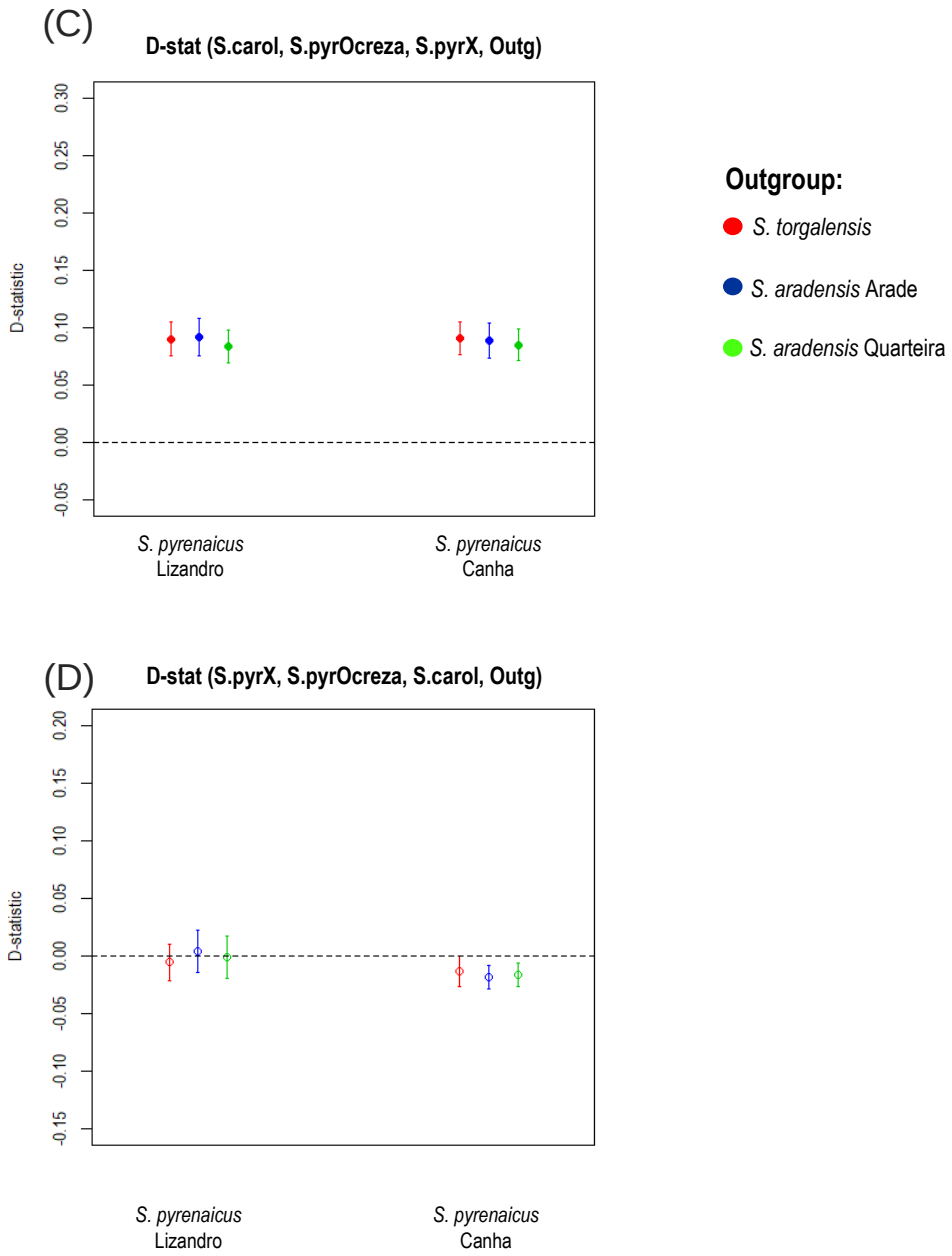


Figure 3.7 (cont.) – Results of the D statistic calculated for the different scenarios in Fig. 2.1. For each topology (A to D), the results are presented according to the northern *S. pyrenaicus* population (S. pyrX) used. In A and B, two plots are presented, one for each southern *S. pyrenaicus* population used: Almargem (S.pyrAlmargem) and Quarteira (S.pyrQuarteira). "S.carol" stands for *S. carolitertii*, S.pyrOcreza stands for *S. pyrenaicus* Ocreza and "Outg" for outgroup. The result using each outgroup is coded with a different colour. Full dots represent significant D values ($p<0.01$).

especially when the outgroup was *S. aradensis* (Figure 3.7 A.2). Nonetheless, across all combinations tested, the D-statistic was always positive reflecting that there were significantly more sites with the pattern ABBA, that is, where the northern *S. pyrenaicus* populations (P2) shares the same allele with *S. carolitertii* (P3). This can be interpreted as a sign of introgression between P2 and P3, or due to a more recent shared ancestry between P2 and P3.

To test the hypothesis that *S. carolitertii* and the northern *S. pyrenaicus* share a more recent ancestry, we tested a species tree where *S. carolitertii* (P1) and the northern *S. pyrenaicus* (P2) are treated as sister species, with the southern *S. pyrenaicus* acting as a potential source of introgressed alleles (P3) (Figure 2.2 B). In this case, most of combinations of sampling locations resulted in D-sta-

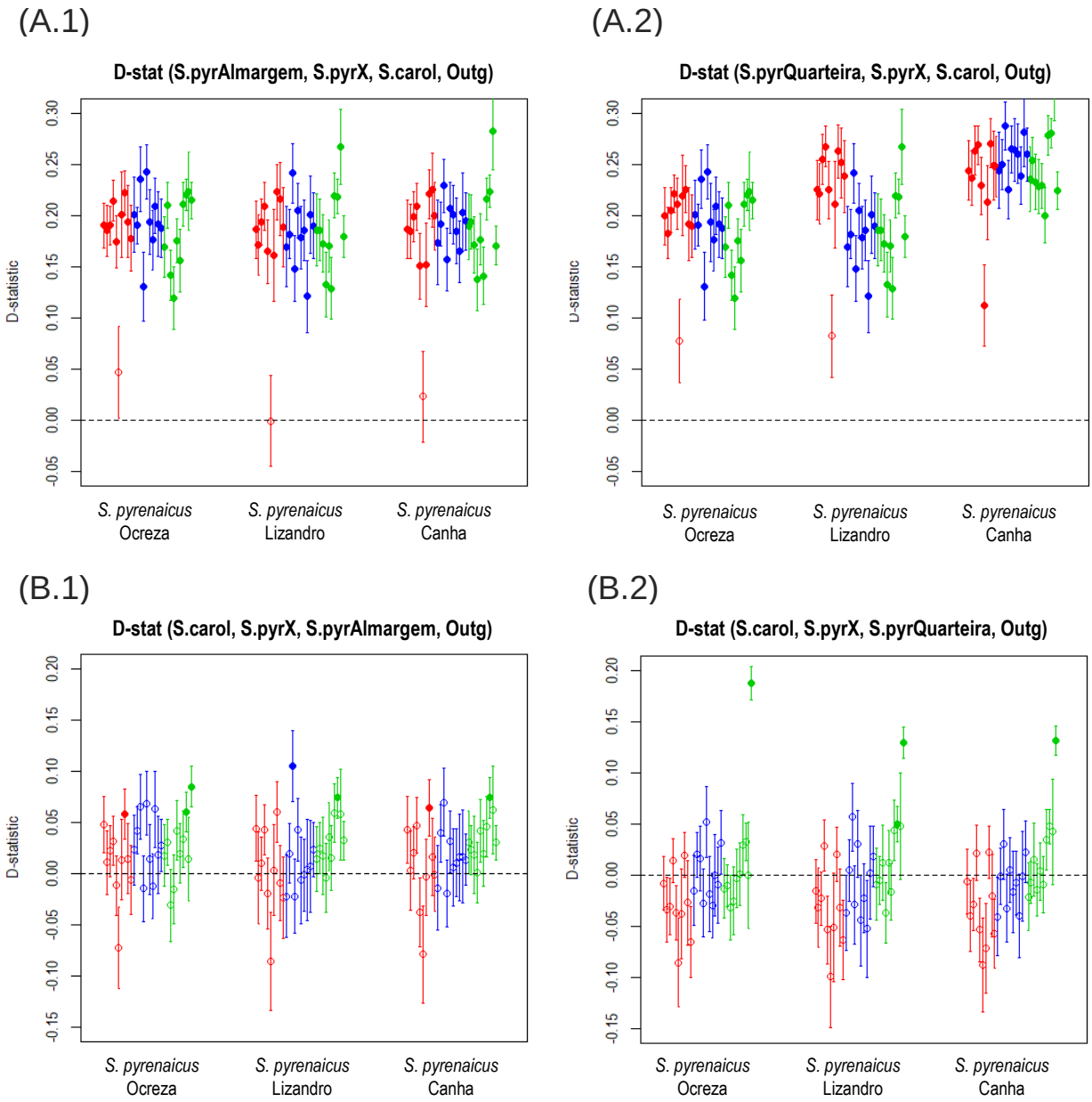


Figure 3.8 - Results of the D statistic calculated per individual for the different scenarios in Fig. 2.1. For each topology (A to D), the results fare presented according to the northern *S. pyrenaicus* population (S. pyrX) used. Each point represents one individual. In A and B, two plots are presented, one for each southern *S. pyrenaicus* population used: Almargem (S.pyrAlmargem) and Quarteira (S.pyrQuarteira). "S.carol" stands for *S. carolitertii*, S.pyrOcreza stands for *S. pyrenaicus* Ocreza and "Outg" for outgroup. The result using each outgroup is coded with a different colour. Full dots represent significant D values ($p<0.01$).

tistic values not significantly different from zero (Figure 3.7 B). This indicates that the southern *S. pyrenaicus* is equally distant from *S. carolitertii* and the northern *S. pyrenaicus*. However, we found some exceptions. In particular, in many combinations where P2 is *S. pyrenaicus* Canha, D was slightly positive and significantly different from zero, indicating that *S. pyrenaicus* Canha was closer to the southern *S. pyrenaicus*. We note, however, that this pattern was not found when P3 was *S. pyrenaicus* Almargem and the outgroup was either of the *S. aradensis* populations. Overall, these results are in agreement with those from the PCA and sNMF and with the species tree inferred with TreeMix, suggesting that rather than a scenario of introgression from *S. carolitertii* into northern *S. pyrenaicus*, they share a more recent common ancestor.

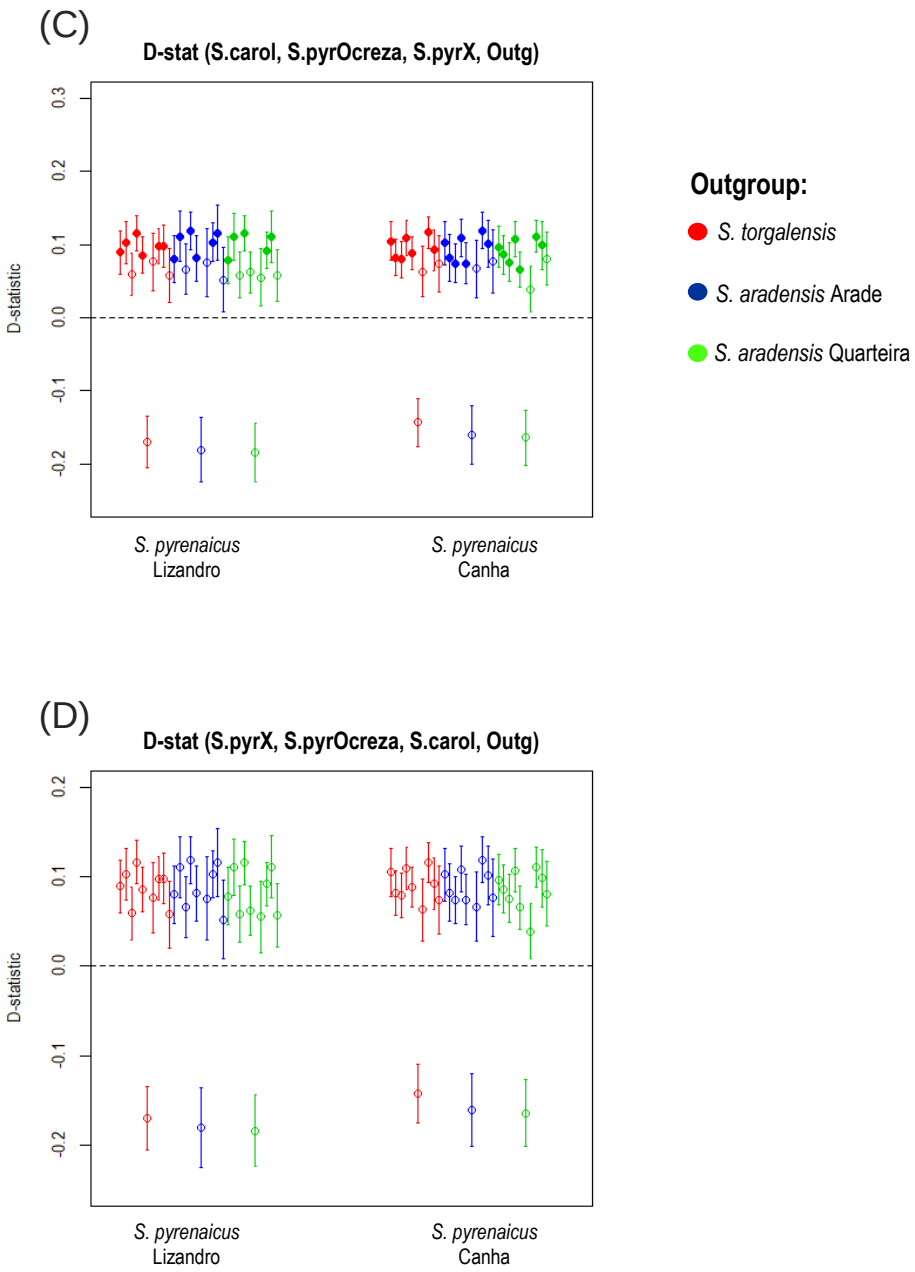


Figure 3.8 (cont.) - Results of the D statistic calculated per individual for the different scenarios in Fig. 2.1. For each topology (A to D), the results fare presented according to the northern *S. pyrenaicus* population (S. pyrX) used. Each point represents one individual. In A and B, two plots are presented, one for each southern *S. pyrenaicus* population used: Almargem (S.pyrAlmargem) and Quarteira (S.pyrQuarteira). "S.carol" stands for *S. carolitertii*, S.pyrOcreza stands for *S. pyrenaicus* Ocreza and "Outg" for outgroup. The result using each outgroup is coded with a different colour. Full dots represent significant D values ($p<0.01$).

If *S. carolitertii* diverged at different times from the northern *S. pyrenaicus* populations, or if introgression occurred after divergence, we would expect differences in D-statistics among the northern *S. pyrenaicus*. To investigate the possibility that the northern most sampling location of *S. pyrenaicus* (Ocreza) is closer to *S. carolitertii* than the other northern *S. pyrenaicus* locations, we computed D-statistics according to the species tree in Figure 2.1 C. The estimated D-values were always significantly positive (Figure 3.7 C), which indicates that *S. pyrenaicus* Ocreza shares more alleles with the other northern *S. pyrenaicus* than with *S. carolitertii*. Contrarily, when the sister populations are both from the northern area of *S. pyrenaicus* distribution and P3 is *S. carolitertii* (Figure 2.2 D), D is never significantly different from zero (Figure 3.7 D). These results indicate that *S. pyrenaicus* Ocreza is not closer

to *S. carolitertii*, suggesting that all northern *S. pyrenaicus* populations share similar numbers of alleles with *S. carolitertii*. This is consistent with the species tree inferred with TreeMix, where the species tree shows that all northern *S. pyrenaicus* have a common ancestor that diverged from *S. carolitertii* after the divergence of the southern *S. pyrenaicus* from the lineage that originated *S. carolitertii* and the northern *S. pyrenaicus* (Figure 3.6). However, a scenario of introgression between *S. carolitertii* and the ancestor of the northern *S. pyrenaicus* (i.e. prior to the divergence of the different northern *S. pyrenaicus* populations) could also lead to the same results.

If there were recent introgression events, we would expect to find differences in the D-statistic value among different individuals from a given population. To detect evidence of such relatively recent introgression between species, we computed the D-statistic by individual. The exact value of the D-statistic and p-values can be found on Supplementary Table S12, where significant p-values are highlighted in grey shading.

For the first topology (Figure 2.2 A), the results for all individuals were consistent with those obtained when the statistic was calculated per population. All individuals from P2 indicated positive and significant D-statistic values, irrespective of the combination of populations P1 and P2 (Figure 3.8 A). Only one individual from *S. pyrenaicus* Ocreza was the exception, constantly exhibiting a D-statistic value that is not significantly different from zero (Figure 3.8 A). This outlier individual shows a high percentage of missing data (≈59.19), and thus we interpret this as an artefact due to a lower number of sites in that individual.

For the second topology (Figure 2.2 B), where we investigated the possibility that *S. carolitertii* and the northern *S. pyrenaicus* share a more recent common ancestor, most individuals display a value of D that is not significantly different from zero, in line with the results obtained per population (Figure 3.8 B). Despite that, a few individuals show a significant positive D value. Overall, for the first two topologies we find no considerable variation between individuals from the same sampling location – in Figure 3.8 A the values of D were always significantly positive and in Figure 3.8 B the values of D were not different from zero, indicating that the introgression is not ongoing and probably not very recent.

In the third topology (Figure 2.2 C) most individuals show a significant positive D-value (Figure 3.8C), as obtained per population, suggesting that *S. pyrenaicus* Ocreza is not closer to *S. carolitertii* than the other northern *S. pyrenaicus* locations. However, some exceptions occur, as some individuals show a D value that is not significantly different from zero. The fact that for the fourth topology all individuals have a D-statistic not significantly different from zero (Figure 3.8 D) indicates that the northern *S. pyrenaicus* populations are not differentially distant from *S. carolitertii* (Figure 3.8 D). Therefore, considering the individual results and the ones from Figure 3.7 D, there is no sign of more recent divergence (or stronger introgression) of *S. carolitertii* and *S. pyrenaicus* Ocreza than with the other two northern *S. pyrenaicus* populations (Lizandro and Canha).

Demographic modelling of divergence of *S. carolitertii* and *S. pyrenaicus*

The model with admixture achieved a higher likelihood than the model without admixture (Table 3.4), suggesting that the northern *S. pyrenaicus* received a contribution from both *S. carolitertii* and the southern *S. pyrenaicus*. The estimated parameters are relative to the unknown reference effective size of *S. carolitertii*, and thus Table 3.5 shows the estimated parameter values according to different reference N_e values. Most parameter values were similar across models, and suggest similar population sizes in the three populations, large ancestral sizes, and a relative recent split of the northern *S. pyrenaicus* (approximately 0.15-0.17 of the time split from ancestral). Interestingly, under the best model (admixture), we estimate that at the time of split the northern *S. pyrenaicus* received a contribution

Table 3.4 - Model comparison of estimated likelihood values obtained with *fastsimcoal2*.

	#params[a]	Estimated log10(Likelihood) [b]	Difference to max likelihood[c]
Admixture	8	-14,545	-238
No Admixture	8	-14,560	-254

[a] #param: number of parameters.
[b] Estimated log10(likelihood): estimated likelihood values in log10 scale.
[c] Difference to max likelihood: difference in log10 units between the estimated likelihood and the maximum likelihood if there was a perfect fit to the observed site frequency spectrum. The closer to zero (less negative values), the better the fit.

Table 3.5 - Parameter estimates obtained with *fastsimcoal2* for the two tested models, scaled with different values of the reference effective size. Note that the estimated effective sizes and times of events are estimated relative to the unknown effective size of *S. carolitertii* (reference Ne), and thus parameter estimates are given assuming that the reference *Ne* varies between 10,000 and 10^6. The admixture model attained a higher likelihood and assumes that at the time of divergence the northern *S. pyrenaicus* received a contribution (admixture %) from the southern *S. pyrenaicus*.

	Reference effective size 10,000		Reference effective size 100,000		Reference effective size 1,000,000	
	Admixture	**No Admixture**	**Admixture**	**No Admixture**	**Admixture**	**No Admixture**
Contribution from *S. pyrenaicus* South						
Admixture %	14.4	0	14.4	0	14.4	0
Effective sizes						
S. carolitertii	10,000	10,000	100,000	100,000	1,000,000	1,000,000
S. pyrenaicus North	7,088	427,074	70,877	4,270,739	708,767	42,707,387
Bottleneck *S. pyrenaicus* North	NA	31	NA	306	NA	3,063
S. pyrenaicus South	6,048	6,559	60,477	65,593	604,765	655,934
Ancestral *S. carolitertii*	65,055	99,803	650,555	998,027	6,505,549	9,980,267
Ancestral *S. pyrenaicus*	139,312	205,984	1,393,123	2,059,842	13,931,234	20,598,424
Ancestral	3,625	2,966	36,248	29,659	362,478	296,594
Time of events (Mya)						
Divergence of *S. pyrenaicus* North	0.009	0.009	0.093	0.088	0.925	0.884
Divergence from ancestral	0.052	0.058	0.516	0.585	5.161	5.846

of 14.4% from the southern *S. pyrenaicus* and the remaining 85.6% from *S. carolitertii*. We note that these results do not fully agree with the D-statistic tests, since we would expect significant positive D-statistics for tree B (Figure 2.2 B). Indeed, the D values tend to be positive, but are not significant, suggesting that we have more power with the demographic modelling based on the joint SFS than with the D-statistic to detect introgression.

4. Discussion

In this work, our goal was to investigate the relationship between populations of *S. carolitertii*, *S. pyrenaicus*, *S. aradensis* and *S. torgalensis* using genome-wide data (SNPs) obtained through Genotyping by Sequencing. This is the first time such an approach is employed to these species. With the pipeline we developed, we successfully obtained a high-quality set of SNP markers for these four species from GBS data without a reference genome. This dataset allowed us to infer the species tree describing the relationship between these four species. Furthermore, this genome-wide data allowed us to test for past introgression between *S. carolitertii* and *S. pyrenaicus* in the northern part of *S. pyrenaicus* distribution.

Obtention of a high-quality SNP dataset

Genotype by Sequencing protocols are often used to address questions at the intraspecific level, involving several populations of the same species, especially when a reference genome is not available. In studies involving more than one species, due to the higher levels of divergence between sampled individuals, it becomes more difficult to identify genetic variants and call SNPs without a reference. Here, we developed a pipeline that allowed us to deal with the limitations of analysing paired-end GBS data without a reference genome. A key step is the construction of the catalogue, where we chose similar parameters to those used by Paris et al. 2017 for the dataset with the characteristics that more resembled ours. We were careful in our choice of filters so that we only kept high quality SNPs but also minimized missing data as much as possible. Moreover, to deal with the possible effects of missing data, we made sure that methods that could efficiently deal with this issue (TreeMix and D-statistic) were implemented in conjunction with the other analysis. Furthermore, for the demographic modelling we performed a downsampling, such that at each site there was no missing data, in order to obtain the site-frequency spectrum. Finally, given the absence of a reference genome, we could not map the SNPs to investigate if the dataset contained linked sites. To overcome this, all analysis, with the exception of the D-statistic and demographic modelling, due to the lack of sites, were repeated with a smaller dataset of approximately 4,000 SNPs, sampling 1 SNP per 200bp block. Since we uncovered very similar results to the ones obtained with the full dataset, we interpret this as an indication that linked sites are not a major problem in our data.

Species tree of *S. carolitertii*, *S. pyrenaicus*, *S. torgalensis* and *S. aradensis*

Taken together, our results indicate a species tree composed of two main lineages: (i) *S. aradensis* and *S. torgalensis* and (ii) *S. carolitertii* and *S. pyrenaicus*. This is evidenced by the pairwise F_{ST} results indicating lower levels of differentiation within each lineage than between the two lineages, as well as by the PCA results (Figure 3.3) and the inferred species tree (Figure 3.5). This is in agreement with phylogenies previously obtained for cytochrome b (Brito et al. 1997; Sanjur et al. 2003; Doadrio and Carmona 2006; Mesquita et al. 2007; Perea et al. 2010) and nuclear genes (Almada and Sousa-Santos 2010; Waap et al. 2011). The divergence between the two main lineages has been dated to have taken place during the Miocene (Mesquita et al. 2007) and recent estimates with mitochondrial and nuclear genes put their divergence approximately at 14 Million years ago (Mya) (Coelho et al, in prep). At that point, the configuration of the river systems was very different from today, characterized by many endorheic basins (basins that did not flow to the ocean). The Tagus was composed of several endorheic lakes and the isolation of one of them, the Lower Tagus (approximately in the current location of the Tagus and Sado river mouths) has been suggested as a possible explanation for the isolation of

the ancestor of *S. aradensis* and *S. torgalensis*, which then migrated south, to their current distributions (Sousa-Santos et al. 2007, Coelho et al., in prep).

Concerning the southwestern species (*S. aradensis* and *S. torgalensis)*, their differentiation into two distinct species has been proposed as a result of the uplift of the Caldeirão mountains, in the south of Portugal, which contributed to the isolation of the ancestors of *S. torgalensis* and *S. aradensis* in the Mira and Arade river basins respectively (Mesquita et al. 2005), with the most recent estimates of their divergence pointing to 4 Mya (Coelho et al, in prep). Regarding *S. aradensis*, we found no evidence of significant genetic differentiation between the two analysed sampling locations of this species (Table 3.2 and Figures 3.3 and 3.4). Although a previous study describes very fragmented populations of this species along several small river basins, we note that the lowest genetic differentiation found between populations of different drainages was precisely between Arade and Quarteira (Mesquita et al. 2005). Moreover, that study focused only on *S. aradensis*, and therefore aimed at detecting fine structure within the species, by having a dense sampling across the species range, while ours is looking at a broader scale, using data from four species to infer a species tree. Interestingly, in Quarteira both *S. aradensis* and *S. pyrenaicus* are found, so this is the only sampling location in our dataset where two species are found in sympatry. Although our PCA and sNMF analysis (Figures 3.3 and 3.4) showed no indication for introgression between *S. aradensis* and *S. pyrenaicus* in Quarteira, in the future similar methods to the ones we used for *S. carolitertii* and *S. pyrenaicus* could be employed to further explore this question. Here, due to the lack of a suitable outgroup, we could not perform the D-statistic tests to investigate the possibility of admixture between *S. aradensis* and *S. pyrenaicus* in Quarteira.

For *S. torgalensis*, we found that the expected heterozygosity was higher than the observed heterozygosity (Figure 3.2), which is in agreement with previous results based on microsatellite markers (Henriques et al. 2010). As previously suggested by Henriques et al. 2010, this might be a result of complex dispersal patterns across the Mira river, which is characterized by very pronounced changes in the hydrological regime according to the seasons (floods in the winter and drought in the summer).

Introgression between *S. carolitertii* and *S. pyrenaicus*

For the second lineage, comprising *S. carolitertii* and *S. pyrenaicus*, previous studies suggested the possibility of introgression to explain incongruences found between nuclear and mitochondrial markers. Our results indicate that *S. pyrenaicus* is paraphyletic along its distribution: *S. pyrenaicus* from the north (Ocreza, Lizandro and Canha) are genetically closer to *S. carolitertii* than to the southern *S. pyrenaicus* (Almargem, Guadiana, Quarteira). This contradicts the mitochondrial phylogenies where, with few exceptions (e.g Zêzere river), all *S. pyrenaicus* populations formed a monophyletic group (e.g. Brito et al. 1997) but is in accordance with the work done with a set of candidate nuclear genes, where this paraphyly with respect to *S. carolitertii* was also uncovered using only samples from the right margin of the Tagus basin (Waap et al. 2011). Here, we confirm that the closer genetic proximity of the northern *S. pyrenicus* to *S. carolitertii* (i) is a genome wide pattern and (ii) does not affect only the right margin of the Tagus, geographically closer to *S. carolitertii*. Although we only have *S. carolitertii* samples from the Mondego, we have no reason to suspect the species tree would change significantly with the inclusion of other *S. carolitertii* populations, since no study has ever reported incongruences regarding the monophyly of *S. carolitertii*, either with mitochondrial or nuclear markers (Almada and Sousa-Santos 2010; Waap et al. 2011, Coelho et al., in prep). However, as discussed in more detail below, a wider sampling across *S. carolitertii* range would be required to further confirm this (e.g. Douro and Vouga basins).

Previous studies described *S. carolitertii* as having low genetic diversity, when compared to the other *Squalius* (e.g. Coelho et al. 1995; Brito et al. 1997; Sanjur et al. 2003). Interestingly, we do not

find that to be the case in our genome-wide SNP dataset, as *S. carolitertii* exhibits genetic diversity levels comparable to other species. However, *S. carolitertii* was sampled in the Mondego basin (Ceira river), which based on fewer markers seems to be an exception to the pattern of low genetic diversity. In fact, Brito et al. 1997 found the genetic variability of *S. carolitertii* mtDNA (cytochrome b) in Mondego to be three times higher than in other basins, like Douro and Vouga. Other studies have also found a higher number of mitochondrial (cyt b) (Sousa-Santos et al. 2007; Almada and Sousa-Santos 2010) and nuclear beta-actin gene (Almada and Sousa-Santos 2010) private haplotypes in the Mondego, when compared to other river basins. More recently, a dense sampling covering the majority of *S. carolitertii* distribution, although only based in one mitochondrial marker (cyt b), found the highest levels of genetic diversity within the species to be precisely in the Ceira river (Sousa-Santos et al. 2016). The relatively lower levels of genetic diversity of *S. carolitertti* have been attributed to bottlenecks due to the impact of glaciations during the Pleistocene (Brito et al. 1997). This pattern of higher genetic diversity in the Mondego when compared to rivers further north has also been described for other organisms besides freshwater fish, namely for golden-striped salamanders (*Chioglossa lusitanica*) (Alexandrino et al. 2002). Another hypothesis to explain the higher diversity in Mondego, which we also find, would be that the proximity between the Mondego and some tributaries of the Tagus may have allowed the exchange of alleles due to river captures that might have occurred in the past, increasing the genetic diversity in the Mondego basin due to introgression (Brito et al. 1997; Sousa-Santos et al. 2007).

Based on fossil calibrated trees from seven nuclear genes, the most recent common ancestor of *S. carolitertii* and all *S. pyrenaicus* has been dated to 6Mya, while the differentiation between *S. carolitertii* and the northern *S. pyrenaicus* is more recent (3Mya) (Coelho, et al., in prep). Our PCA and sNMF clustering analyses did not allow to distinguish between northern *S. pyrenaicus* and *S. carolitertii* but could separate these from the southern *S. pyrenaicus* (Figures 3.3 and 3.4). In fact, the PCA seems to reflect major divergence events happening in the evolutionary history of these four species: (i) PC1 separates the two main lineages (*S. carolitertii* + *S. pyrenaicus* from *S. torgalensis* + *S. aradensis*); (ii) PC2 further separates the clade of the southern *S. pyrenaicus* from *S. carolitertii* and the northern *S. pyrenaicus*; (iii) and PC3 separates *S. aradensis* from *S. torgalensis* (Figure 3.3). Our sNMF results suggest four clusters, but interestingly individuals are not assigned with 100% to only one of those clusters (Figure 3.4). Many individuals have >70% assigned to one cluster and the remaining into other clusters. This variation can reflect shared ancestral polymorphism, past introgression or uncertainty due to statistical noise and missing data. Given the pairwise F_{ST}, TreeMix and D-statistic results, we interpret that this variation arises mainly due to ancestral polymorphism. Nevertheless, even though most individuals have some ancestry proportions from more than one cluster, they can be grouped into four clusters. Although sNMF does not separate *S. carolitertii* from northern *S. pyrenaicus*, most individuals from southern *S. pyrenaicus* are assigned to a different cluster (Figure 3.4). The ancestry proportions in some individuals from Almargem and one individual from Guadiana (southern *S. pyrenaicus*) appear to have significant proportions assigned to the cluster of *S. carolitertii* and the northern *S. pyrenaicus*, but we note that such individuals were the ones with higher proportion of missing data (Supplementary Table 4), which pulls ancestry proportions towards an equal assignment into each cluster, as expected if there was no information in the data. Nonetheless, when we applied the TreeMix method which efficiently deals with missing data, the placement of Almargem and Guadiana within the southern *S. pyrenaicus* cluster was fully resolved (Figure 3.6).

Considering the results discussed until this point, the most likely explanation to our genome-wide results seems to be that the correct species tree is *S. carolitertii* and the northern *S. pyrenaicus* sharing a more recent common ancestor after their divergence from the lineage of the southern *S. pyrenaicus*. Indeed, our pairwise F_{ST} results seem to support this topology and show that the southern *S. pyrenaicus* lineage is more differentiated from *S. carolitertii* and the northern *S. pyrenaicus* than these

two lineages are from each other (Table 3.2). Excluding Lizandro, within the northern *S. pyrenaicus*, the pairwise F_{ST} values are lower between the left (Canha) and right (Ocreza) margin of the Tagus than between the northern *S. pyrenaicus* and *S. carolitertii*, despite some overlap. The northern *S. pyrenaicus* from Lizandro show slightly higher differentiation from all the other species and sampling locations than the other northern *S. pyrenaicus*. Coupled with its lower genetic diversity (Figure 3.2), this result may be due to stronger genetic drift due to long term isolation in this small basin, likely associated with a bottleneck when it was colonized, probably from the Tagus, as has been hypothesized for another small basin nearby (Colares) (Sousa-Santos et al. 2007).

The pattern of lower genetic differentiation between the northern *S. pyrenaicus* and *S. carolitertii* than between northern and southern *S. pyrenaicus* could be explained by two different scenarios: (i) *S. carolitertii* and the northern *S. pyrenaicus* share a more recent common ancestor but evolved independently in the absence of gene flow; (ii) the northern *S. pyrenaicus* appear closer to *S. carolitertii* due to extensive introgression between them. The topology of our inferred species tree could, in principle, be explained by both scenarios (Figure 3.6). Our results of the D-statistic help to unravel some important patterns (Figures 3.7 and 3.8). First, the fact that our D values for tree A (Figure 2.2) are consistent regardless of the northern *S. pyrenaicus* used (Ocreza, Lizandro or Canha) indicates that any possible introgression between *S. carolitertii* and *S. pyrenaicus* had to be older than the differentiation of the northern *S. pyrenaicus* in different rivers. In fact, if introgression occurred it had to be older than the isolation of *S. pyrenaicus* in Lizandro, a small basin with no contact with the Tagus basin. The D statistic values are also consistent regardless of the southern *S. pyrenaicus* sample used (Almargem or Quarteira). Furthermore, the fact D-statistic results are consistent regardless of the outgroup species used (*S. aradensis* or *S. torgalensis*) indicates that the two main groups inferred in the species tree (*S. aradensis* + *S. torgalensis* and *S. carolitertii* + S. *pyrenaicus*) evolved independently. Any differences when using different outgroups would mean the outgroups were not fully isolated from *S. pyrenaicus* and *S. carolitertii*, but that is clearly not the case and suggests no major past events of gene flow between species of the two major clades, i.e. between (*S. aradensis*, *S. torgalensis*) and (*S. pyrenaicus* North, *S. pyrenaicus* South, *S. carolitertii*). If we assume that *S. carolitertii* and the northern *S. pyrenaicus* are sister species (Figure 2.2), the values of the D-statistic were almost never significantly different from zero. Combined with the fact that any possible introgression had to be older to the isolation of the fish in different tributaries, this seems to indicate that rather than invoking introgression between *S. carolitertii* and the northern *S. pyrenaicus*, an easier explanation would be that the northern *S. pyrenaicus* share a more recent common ancestor with *S. carolitertii* and evolved independently without gene flow. However, even though not significant, the D-values tend to be positive, suggesting some excess of shared alleles between northern and southern *S. pyrenaicus*. Nonetheless, they do not allow us to distinguish between hypotheses (i) and (ii).

Demographic modelling based on the joint 3 population site frequency spectrum allowed us to test these two different scenarios, one with (admixture) and another without (no admixture) past gene flow (Figure 2.3). Our estimates supported that the most likely model was the admixture model, with gene flow (Figure 2.3 A), with an inferred contribution of ≈14.4% from the southern *S. pyrenaicus* and the remaining 85.6% from *S. carolitertii* at the time of the split of the northern *S. pyrenaicus*. These are very simple models but, nonetheless, indicate that northern *S. pyrenaicus* seems to be a mixture of *S. carolitertii* and the southern *S. pyrenaicus* lineage, with a higher proportion from *S. carolitertii* (≈86%). This could explain why *S. pyrenaicus* from the Tagus and Guadiana cluster together in previously inferred mtDNA phylogenies but seem to group in different clusters on nuclear and genome-wide data. The fact that we infer a relatively small admixture contribution from the southern *S. pyrenaicus* (≈14%) is probably the reason why this introgression was not detected with the D-statistics for the tests performed under the tree topology on Figure 2.2 B.

In sum, the demographic modelling parameter estimates support that the divergence of *S. pyrenaicus* and *S. carolitertii* involved events of introgression, and thus the species tree cannot be simply explained by a bifurcating tree. Our model assumes that the time of the admixture with the southern *S. pyrenaicus* is the same as with *S. carolitertii* (Figure 2.3), which corresponds to a scenario of hybrid speciation. Indeed, our estimates raise the possibility that *S. pyrenaicus* from Tagus drainage is the result of hybridization between the southern Guadiana drainage lineages and *S. carolitertii* lineages, which could have happened during the changes of endorheic paleo-drainage systems. In fact, hybrid speciation has been invoked to explain incongruences between nuclear and mtDNA markers and has been proposed in several instances in freshwater fish (DeMarais et al. 1992; Nolte et al. 2005; Meier, Marques, et al. 2017).

However, we cannot discard a secondary contact scenario where after the three lineages were established, gene flow could have occurred from the southern to the northern *S. pyrenaicus*, during periods of river captures and other changes in the drainage systems, explaining the inferred admixture proportions. In fact, the history of the hydrological basins seems to indicate that connections between the Lower Tagus paleobasin and the Guadiana paleobasin ceased before those between the Upper Tagus and the Douro paleobasins (the last two located in present day Spain, near present day Tagus and Douro river springs, respectively) (Coelho et. al, in prep). In this scenario, however, connections had to be re-established between the Tagus and Guadiana basins for the secondary contact to occur. The possibility that the Tagus and Guadiana basins were connected more recently has been proposed to explain the presence of a common lineage in these two basins for another Iberian endemic cyprinid (*Iberochondrostoma lemmingii*) (Lopes-Cunha et al. 2012). Another possibility is that the introgression of southern *S. pyrenaicus* lineages into the northern *S. pyrenaicus* was not caused by the re-establishment of connections between the Tagus and the Guadiana, but between the Tagus and the Sado. In fact, the Lower Tagus and the paleobasin that originated the Sado (Alvalade paleobasin) could have been connected at a time where *S. pyrenaicus* was already present in the Alvalade paleobasin (Coelho et al., in prep).

Final remarks

In face of the incongruent results between mitochondrial and nuclear markers, previous studies have suggested that populations from the Tagus river basin could correspond to a new taxa (Waap et al. 2011). Overall, our results indicate that the patterns observed in the Tagus are most likely the result of introgression. This opens the door to future work to further understand the patterns of introgression between the lineages of *S. carolitertii* and *S. pyrenaicus* in the Tagus basin. Such studies should also include sampling of two key locations that are missing from our dataset: the Zêzere river (a tributary of the Tagus) and the Sado basin. The Zêzere river has consistently been a source of incongruences in mtDNA phylogenies, with some authors finding that *S. pyrenaicus* from this location cluster with *S. carolitertii* (Brito et al. 1997), while others admit that both *S. carolitertii* and *S. pyrenaicus* can be found in this river (Almada and Sousa-Santos 2010; Sousa-Santos et al. 2016). On the other hand, *S. pyrenaicus* from the Sado, although clustering with the Guadiana individuals, in both mitochondrial and nuclear markers, on the phylogenetic analysis (Brito et al. 1997; Waap et al. 2011), have been described as very differentiated from other southern *S. pyrenaicus* in both mitochondrial (cytb) and one nuclear marker (beta-actin gene) (Sousa-Santos et al. 2007; Almada and Sousa-Santos 2010). Unfortunately, in our study, neither the Zêzere river or the Sado basin could be investigated. Furthermore, we only had two samples from the Guadiana basin, both from the Oeiras stream (a tributary of the Guadiana). In the future, upstream localities in the Guadiana basin, closer to the Tagus, should also be sampled. Finally, a broader sampling of the *S. carolitertii* distribution range would also be necessary to better understand the genetic variability found in the Mondego.

If a broader sampling could be obtained, as well as more SNP markers, the two possible scenarios (hybrid speciation or secondary contact) could be further investigated and possibly distinguished. In the future if we obtained data from the proportion of monomorphic sites in the genome and an estimate for the mean genome-wide mutation rate, the absolute time of the divergence or introgression events could also be estimated with demographic modelling. Nonetheless, despite not allowing to distinguish between these two scenarios, our work shows that in the northern *S. pyrenaicus* genome there is contribution from both *S. carolitertii* and the southern *S. pyrenaicus* which calls upon further investigation to understand the evolutionary history behind this pattern.

References

Alexandrino J, Arntzen JW, Ferrand N. 2002. Nested clade analysis and the genetic evidence for population expansion in the phylogeography of the golden-striped salamander, Chioglossa lusitanica (Amphibia: Urodela). Heredity (Edinb). 88:66–74.

Allendorf FW. 2017. Genetics and the conservation of natural populations: allozymes to genomes. Mol. Ecol. 26:420–430.

Almada V, Sousa-Santos C. 2010. Comparisons of the genetic structure of Squalius populations (Teleostei, Cyprinidae) from rivers with contrasting histories, drainage areas and climatic conditions based on two molecular markers. Mol. Phylogenet. Evol. 57:924–931.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat. Rev. Genet. 17:81–92.

Bagley RK, Sousa VC, Niemiller ML, Linnen CR. 2017. History, geography and host use shape genomewide patterns of genetic variation in the redheaded pine sawfly (Neodiprion lecontei). Mol. Ecol. 26:1022–1044.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3:1–7.

Barluenga M, Stölting KN, Salzburger W, Muschick M, Meyer A. 2006. Sympatric speciation in Nicaraguan crater lake cichlid fish. Nature 439:719–723.

Brito RM, Briolay J, Galtier N, Bouvet Y, Coelho MM. 1997. Phylogenetic Relationships within Genus Leuciscus (Pisces , Cyprinidae) in Portuguese Fresh Waters , Based on Mitochondrial DNA Cytochrome b Sequences. Mol. Phylogenet. Evol. 8:435–442.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013. Stacks: An analysis tool set for population genomics. Mol. Ecol. 22:3124–3140.

Coelho MM, Bogutskaya NG, Rodrigues JA, Collares-Pereira MJ. 1998. Leuciscus torgalensis, and L. aradensis, two new cyprinids for Portuguese fresh waters. J. Fish Biol. 52:937–950.

Coelho MM, Brito RM, Pacheco TR, Figueiredo D, Pires AM. 1995. Genetic variation and divergence of Leuciscus pyrenaicus and L. carolitertii (Pisces, Cyprinidae). J. Fish Biol. 47:243–258.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

Dasmahapatra KK, Walters JR, Briscoe AD, Davey JW, Whibley A, Nadeau NJ, Zimin A V., Hughes DST, Ferguson LC, Martin SH, et al. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature 487:94–98.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12:499–510.

DeMarais BD, Dowling TE, Marsh PC, Douglas ME, Minckley WL. 1992. Origin of Gila seminuda (Teleostei : Cyprinidae) through introgressive hybridization : Implications for evolution and conservation. Evolution (N. Y). 89:2747–2751.

Doadrio I. 1987. Leuciscus carolitertii n. sp. from the Iberian Peninsula. Senckenb. Biol. 68:301–309.

Doadrio I, Carmona JA. 2003. Testing freshwater Lago Mare dispersal theory on the phylogeny relationships of iberian cyprinid genera Chondrostoma and Squalius (Cypriniformes, Cyprinidae). Graellsia 59:457–473.

Doadrio I, Carmona JA. 2006. Phylogenetic overview of the genus Squalius (Actinopterygii, Cyprinidae) in the Iberian Peninsula, with a description of two new species. Cybium 30:199–214.

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. Mol. Biol. Evol. 28:2239–2252.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:1–10.

Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32:3047–3048.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013. Robust Demographic Inference from Genomic and SNP Data. PLoS Genet. 9.

Frichot E, François O. 2015. LEA: An R package for landscape and ecological association studies. Methods Ecol. Evol. 6:925–929.

Frichot E, Mathieu F, Trouillon T, Bouchard G, François O. 2014. Fast and efficient estimation of individual ancestry coefficients. Genetics 196:973–983.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152.

Gagnaire PA, Pavey SA, Normandeau E, Bernatchez L. 2013. The genetic architecture of reproductive isolation during speciation-with-gene-flow in lake whitefish species pairs assessed by rad sequencing. Evolution (N. Y). 67:2483–2497.

Gante HF, Matschiner M, Malmstrøm M, Jakobsen KS, Jentoft S, Salzburger W. 2016. Genomics of speciation and introgression in Princess cichlid fishes from Lake Tanganyika. Mol. Ecol. 25:6143–6161.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. rXiv preprint arXiv:1207.3907 [q-bio.GN]

Henriques R, Sousa V, Coelho MM. 2010. Migration patterns counteract seasonal isolation of Squalius torgalensis, a critically endangered freshwater fish inhabiting a typical Circum-Mediterranean small drainage. Conserv. Genet. 11:1859–1870.

Hey J, Machado CA. 2003. The study of structured populations - New hope for a difficult and divided science. Nat. Rev. Genet. 4:535–543.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. 2010. Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags.Begun DJ, editor. PLoS Genet. 6:e1000862.

Hohenlohe PA, Day MD, Amish SJ, Miller MR, Kamps-Hughes N, Boyer MC, Muhlfeld CC, Allendorf FW, Johnson EA, Luikart G. 2013. Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. Mol. Ecol. 22:3002–3013.

Hudson RR, Slatkint M, Maddison WP. 1992. Estimation of Levels of Gene Flow From DNA Sequence Data. Genetics 589:583–589.

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484:55–61.

Jones JC, Fan S, Franchini P, Schartl M, Meyer A. 2013. The evolutionary history of Xiphophorus fish and their sexually selected sword: A genome-wide approach using restriction site-associated DNA sequencing. Mol. Ecol. 22:2986–3001.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN]

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Li W, Godzik A. 2006. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659.

Lopes-Cunha M, Aboim MA, Mesquita N, Alves MJ, Doadrio I, Coelho MM. 2012. Population genetic structure in the Iberian cyprinid fish Iberochondrostoma lemmingii (Steindachner, 1866): Disentangling species fragmentation and colonization processes. Biol. J. Linn. Soc. 105:559–572.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. 2003. The power and promise of population genomics: from genotyping to genome typing. Nat. Rev. Genet. 4:981–994.

McManus KF, Kelley JL, Song S, Veeramah KR, Woerner AE, Stevison LS, Ryder OA, Project GAG, Kidd JM, Wall JD, et al. 2015. Inference of gorilla demographic and selective history from whole-genome sequence data. Mol. Biol. Evol. 32:600–612.

Meier JI, Marques DA, Mwaiko S, Wagner CE, Excoffier L, Seehausen O. 2017. Ancient hybridization fuels rapid cichlid fish adaptive radiations. Nat. Commun. 8:1–11.

Meier JI, Sousa VC, Marques DA, Selz OM, Wagner CE, Excoffier L, Seehausen O. 2017. Demographic modelling with whole-genome data reveals parallel origin of similar Pundamilia cichlid species after hybridization. Mol. Ecol. 26:123–141.

Mesquita N, Cunha C, Carvalho GR, Coelho MM. 2007. Comparative phylogeography of endemic cyprinids in the south-west Iberian Peninsula: Evidence for a new ichthyogeographic area. J. Fish Biol. 71:45–75.

Mesquita N, Hänfling B, Carvalho GR, Coelho MM. 2005. Phylogeography of the cyprinid Squalius aradensis and implications for conservation of the endemic freshwater fauna of southern Portugal. Mol. Ecol. 14:1939–1954.

Miller SA, Dykes DD, Polesky HF. 1988. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res. 16:1215–1215.

Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA. 2013. Genotyping-by-sequencing in ecological and conservation genomics. Mol. Ecol. 22:2841–2847.

Nelson JS, Grande TC, Wilson MVH. 2016. Fishes of the World. Hoboken, NJ, USA: John Wiley & Sons, Inc

Nichols R. 2001. Gene trees and species trees are not the same. Trends Ecol. Evol. 16:358–364.

Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. 12:443–451.

Nolte AW, Freyhof J, Stemshorn KC, Tautz D. 2005. An invasive lineage of sculpins, Cottus sp. (Pisces, Teleostei) in the Rhine with new habitat adaptations has originated from hybridization between old phylogeographic groups. Proc. R. Soc. B Biol. Sci. 272:2379–2387.

Paris JR, Stevens JR, Catchen JM. 2017. Lost in parameter space: a road map for stacks. Methods Ecol. Evol. 8:1360–1373.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. PLoS Genet. 2:2074–2093.

Perea S, Böhme M, Zupancic P, Freyhof J, Sanda R, Ozuluğ M, Abdoli A, Doadrio I. 2010. Phylogenetic relationships and biogeographical patterns in Circum-Mediterranean subfamily Leuciscinae

(Teleostei, Cyprinidae) inferred from both mitochondrial and nuclear data. BMC Evol. Biol. 10:265.

Perea S, Cobo-Simon M, Doadrio I. 2016. Cenozoic tectonic and climatic events in southern Iberian Peninsula: Implications for the evolutionary history of freshwater fish of the genus Squalius (Actinopterygii, Cyprinidae). Mol. Phylogenet. Evol. 97:155–169.

Pfenninger M, Patel S, Arias-Rodriguez L, Feldmeyer B, Riesch R, Plath M. 2015. Unique evolutionary trajectories in repeated adaptation to hydrogen sulphide-toxic habitats of a neotropical fish (Poecilia mexicana). Mol. Ecol. 24:5446–5459.

Pickrell JK, Pritchard JK. 2012. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLoS Genet. 8.

Redenbach Z, Taylor EB. 2002. Evidence for historical introgression along a contact zone between two species of char (Pisces: Salmonidae) in northwestern North America. Evolution (N. Y). 56:1021–1035.

Sanjur OI, Carmona JA, Doadrio I. 2003. Evolutionary and biogeographical patterns within Iberian populations of the genus Squalius inferred from molecular data. Mol. Phylogenet. Evol. 29:20–30.

Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe PA, Peichel CL, Saetre G-P, Bank C, Brännström Å, et al. 2014. Genomics and the origin of species. Nat. Rev. Genet. 15:176–192.

Seehausen O, Wagner CE. 2014. Speciation in Freshwater Fishes. Annu. Rev. Ecol. Evol. Syst. 45:621–651.

Sousa-Santos C, Collares-Pereira MJ, Almada V. 2007. Reading the history of a hybrid fish complex from its molecular record. Mol. Phylogenet. Evol. 45:981–996.

Sousa-Santos C, Robalo JI, Pereira AM, Branco P, Santos JM, Ferreira MT, Sousa M, Doadrio I. 2016. Broad-scale sampling of primary freshwater fish populations reveals the role of intrinsic traits, inter-basin connectivity, drainage area and latitude on shaping contemporary patterns of genetic diversity. PeerJ 4:e1694.

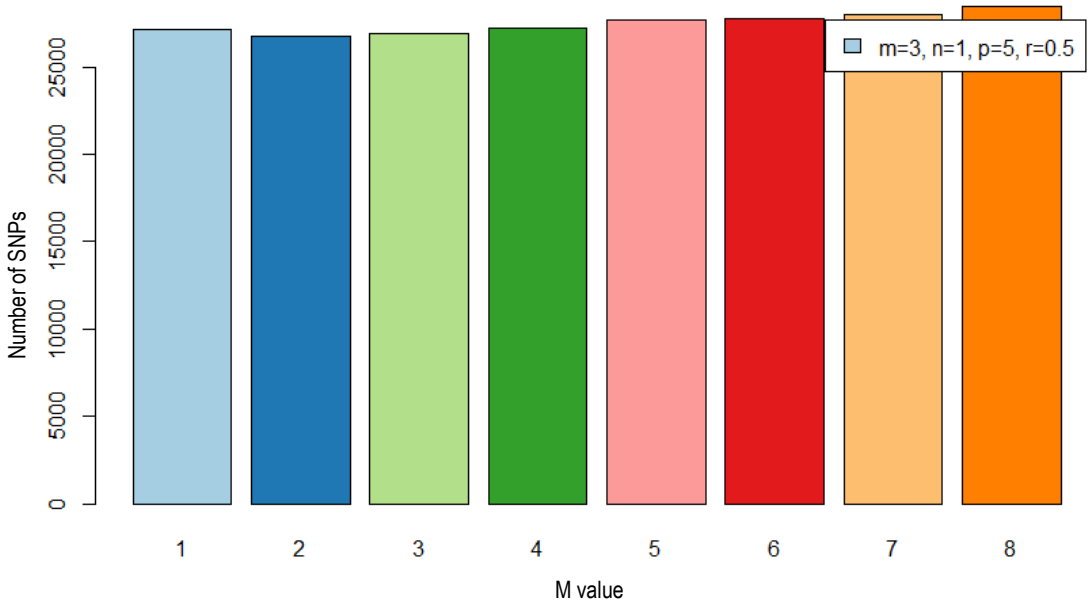
Sousa V, Hey J. 2013. Understanding the origin of species with genome-scale data: modelling gene flow. Nat. Rev. Genet. 14:404–414.

Terekhanova N V., Logacheva MD, Penin AA, Neretina T V., Barmintseva AE, Bazykin GA, Kondrashov AS, Mugue NS. 2014. Fast Evolution from Precast Bricks: Genomics of Young Freshwater Populations of Threespine Stickleback Gasterosteus aculeatus. PLoS Genet. 10.

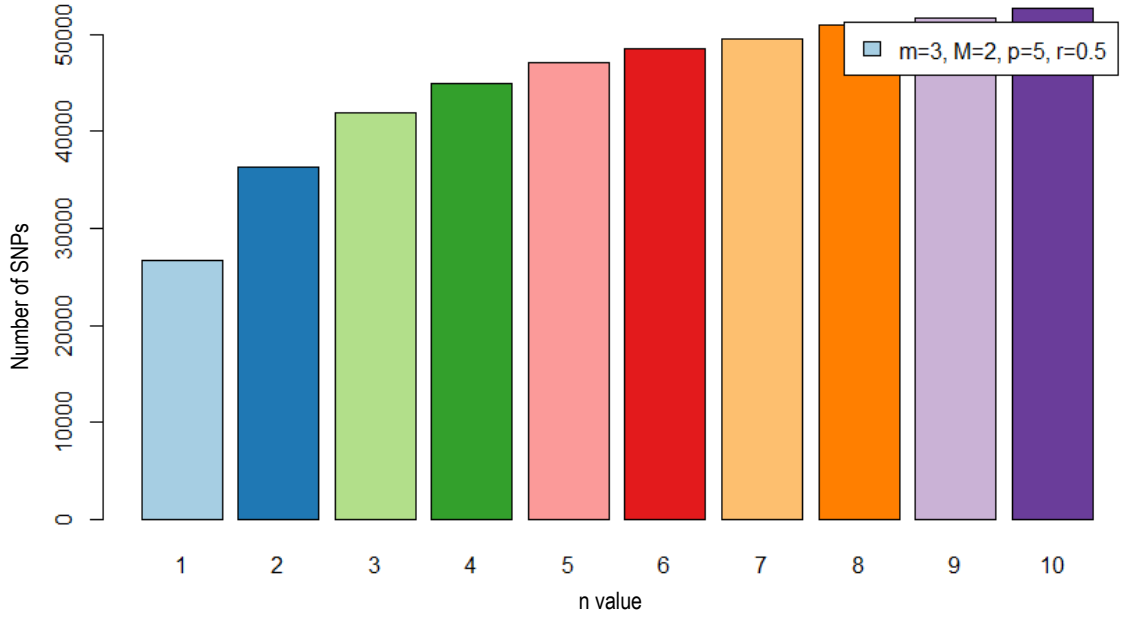
Waap S, Amaral AR, Gomes B, Coelho MM. 2011. Multi-locus species tree of the chub genus Squalius (Leuciscinae: Cyprinidae) from western Iberia: New insights into its evolutionary history. Genetica 139:1009–1018.

Wolf JBW, Ellegren H. 2016. Making sense of genomic islands of differentiation in light of speciation. Nat. Rev. Genet. 18:87–100.

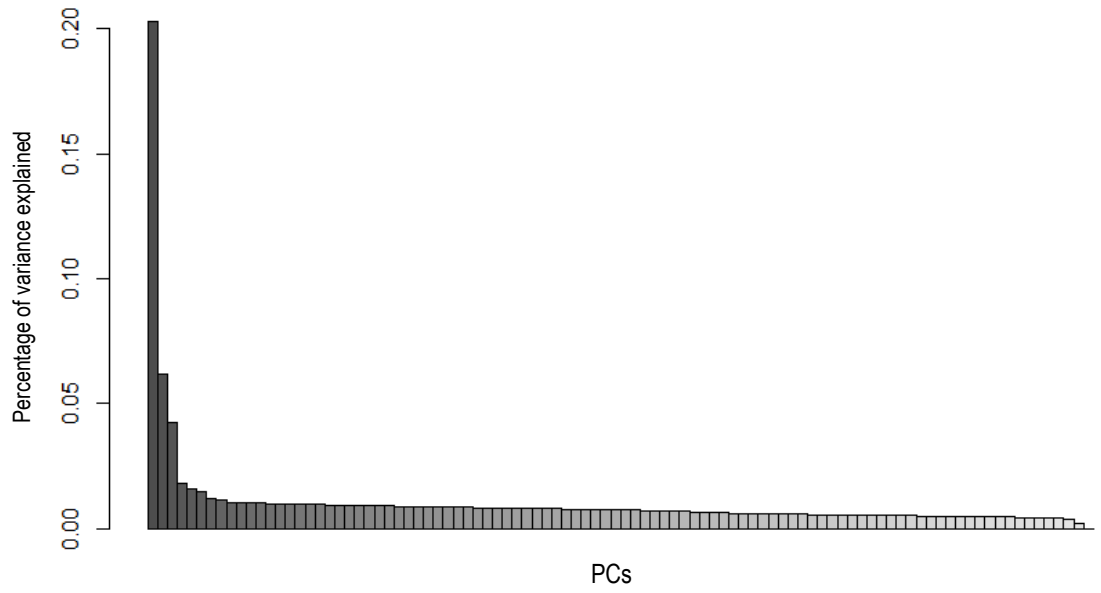
Supplementary Material



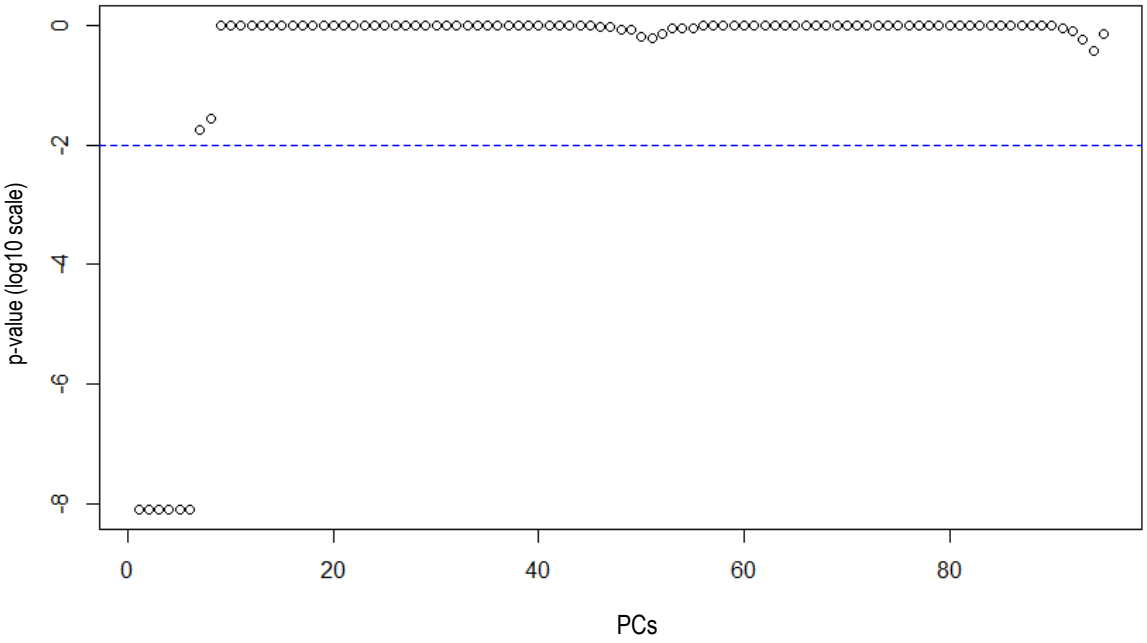
Supplementary Figure S1 – Number of SNPs obtained with different M values. The number of mismatched allowed between sequences from the same individual (*M*) was varied from 1 to 8 while all the other parameters were kept fixed (*m*=3, *n*=1) and SNPs were required to be present in all populations (*p*=5) in 50% of the individuals (*r*=0.5).



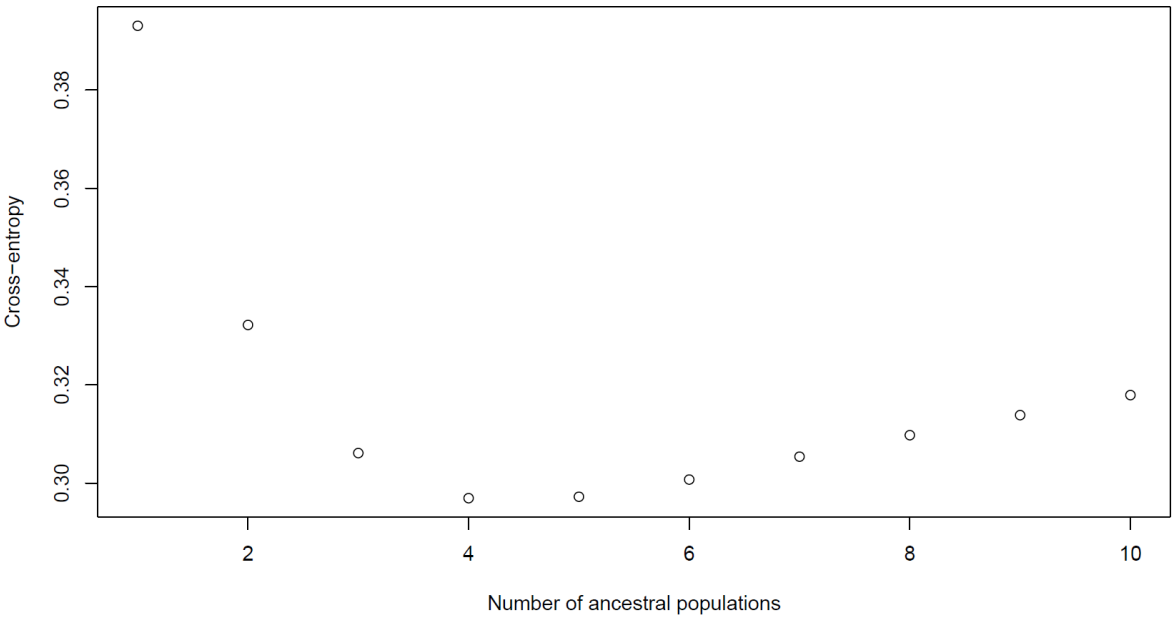
Supplementary Figure S2 – Number of SNPs obtained with different values of n. The number of mismatched allowed between sequences from different individuals (*n*) was varied from 1 to 10 while all the other parameters were kept unchanged (*m*=3, *M*=2) and SNPs were required to be present in all populations (*p*=5) in 50% of the individuals (*r*=0.5).



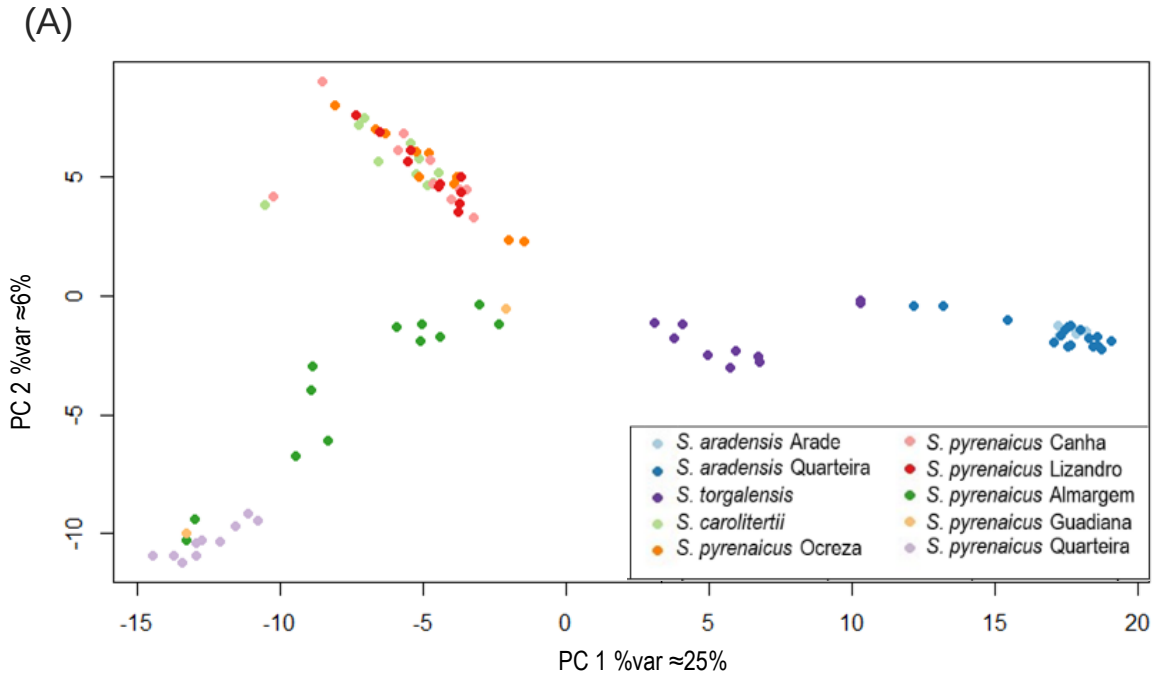
Supplementary Figure S3 – Percentage of variance explained by each principal component (PC) on the Principal Components Analysis (PCA) (Figure 3.3). Together, the first tree components explain ≈30% of the variance.



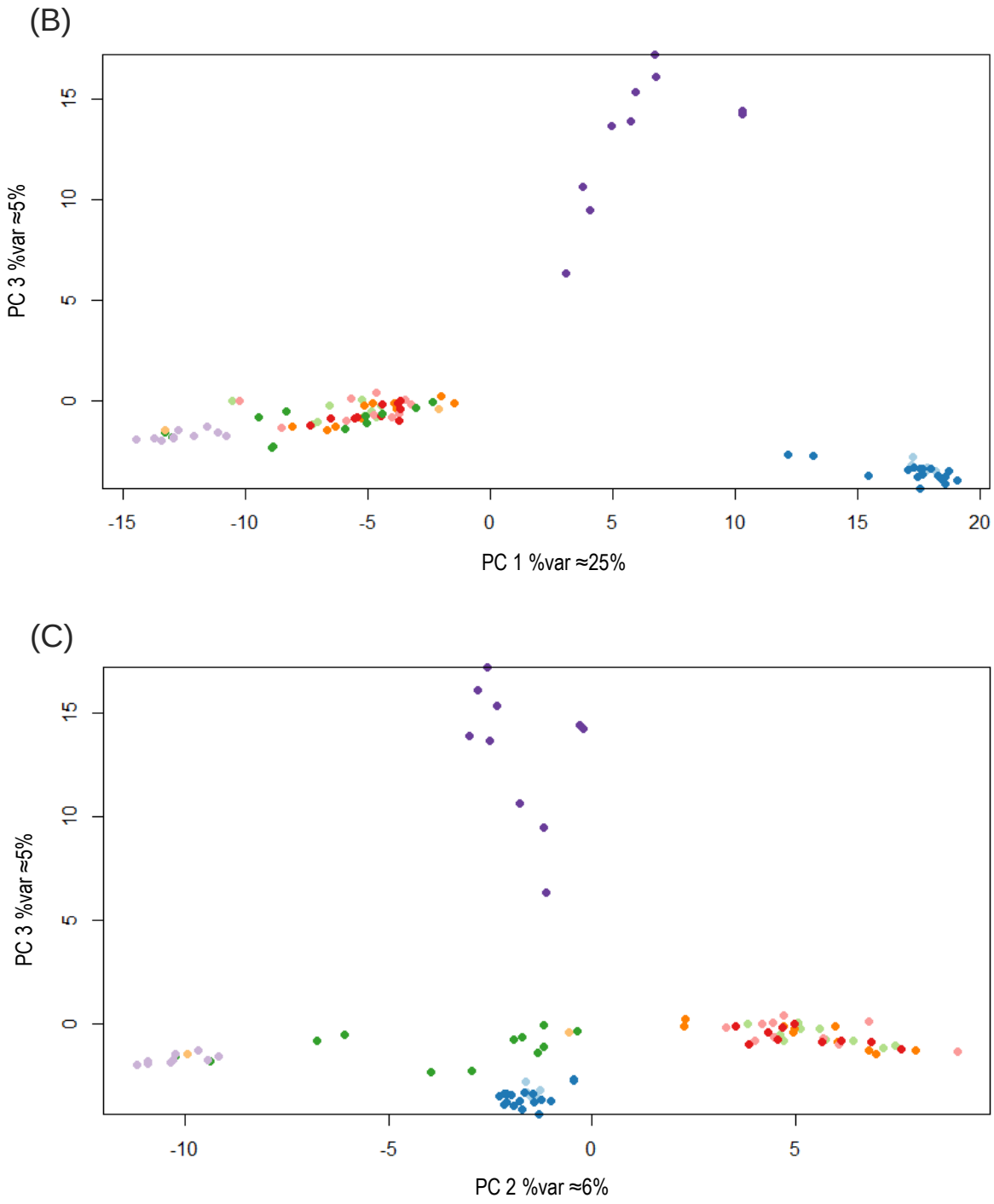
Supplementary Figure S4 – p-values of principal components on the PCA (p<0.01). The y-axis shows the p-values of each principal component (x-axis) in log10 scale. Significance of each principal component was assessed with the Tracy-Widom test. The dashed blue line represent the significance level of 0.01 (log10(0.01)=-2). Principal components below the blue line are considered significant.



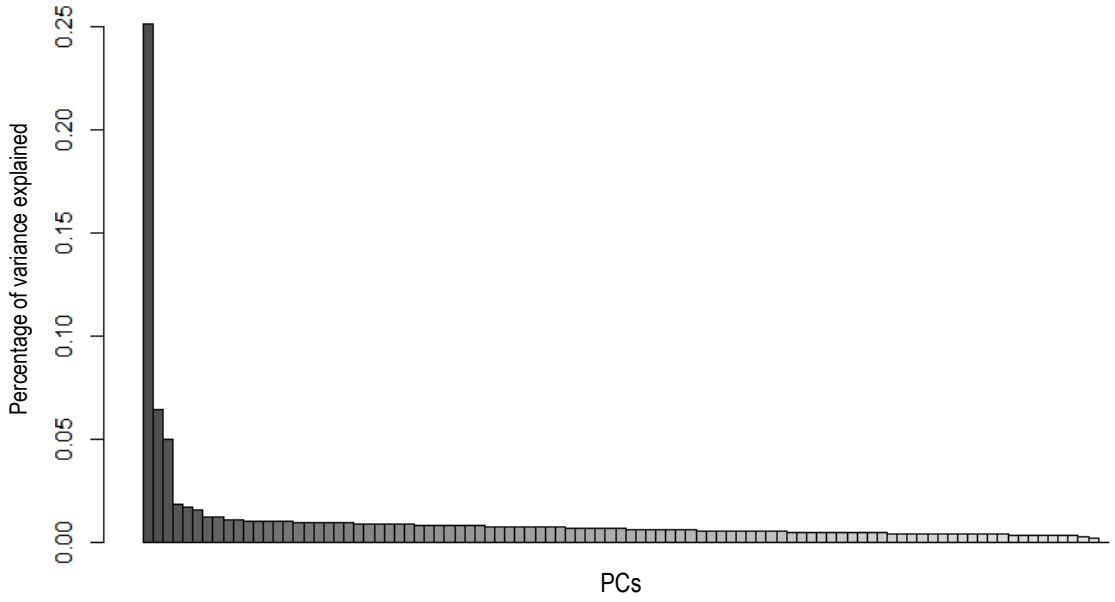
Supplementary Figure S5 – Cross-entropy for each number of K ancestral populations inferred with sNMF. The lowest cross-entropy value indicates the number of ancestral populations that best describes the data.



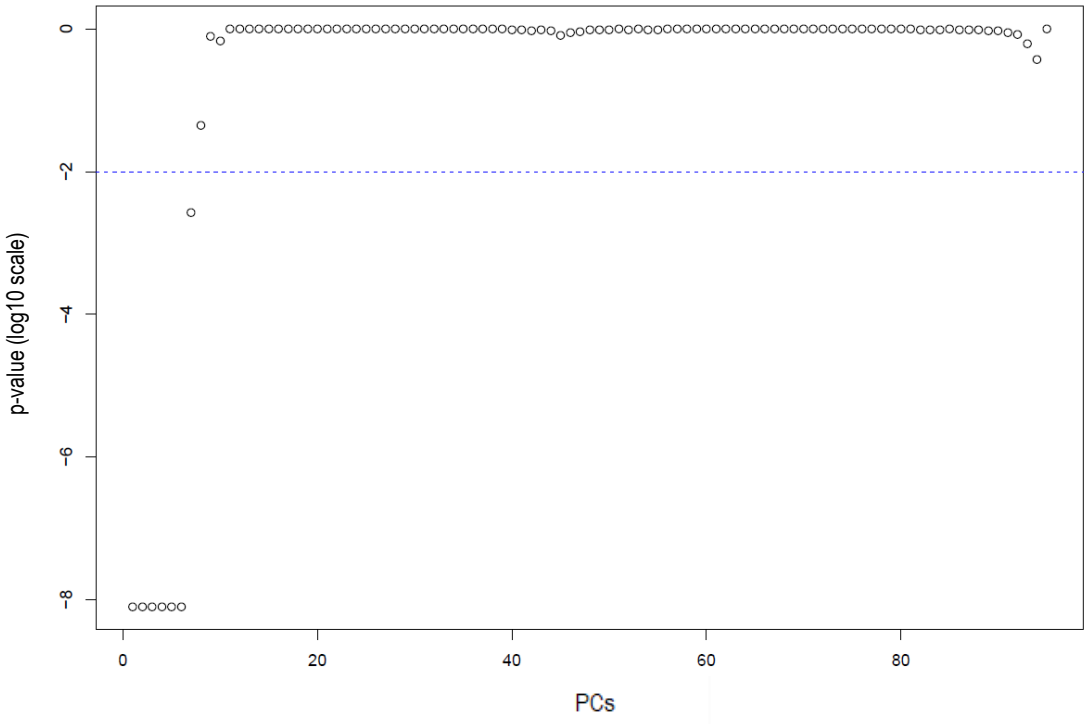
Supplementary Figure S6 – Results for the first tree components of the PCA performed on the dataset with only one SNP per block: (A) PC1 and PC2; (B) PC1 and PC3; (C) PC2 and PC3. Each point corresponds to one individual. The clusters formed are the same obtained with the larger dataset (Fig. 3.2). The percentage of the variance explained by the first principal component slightly increases when compared to Fig. 3.2.



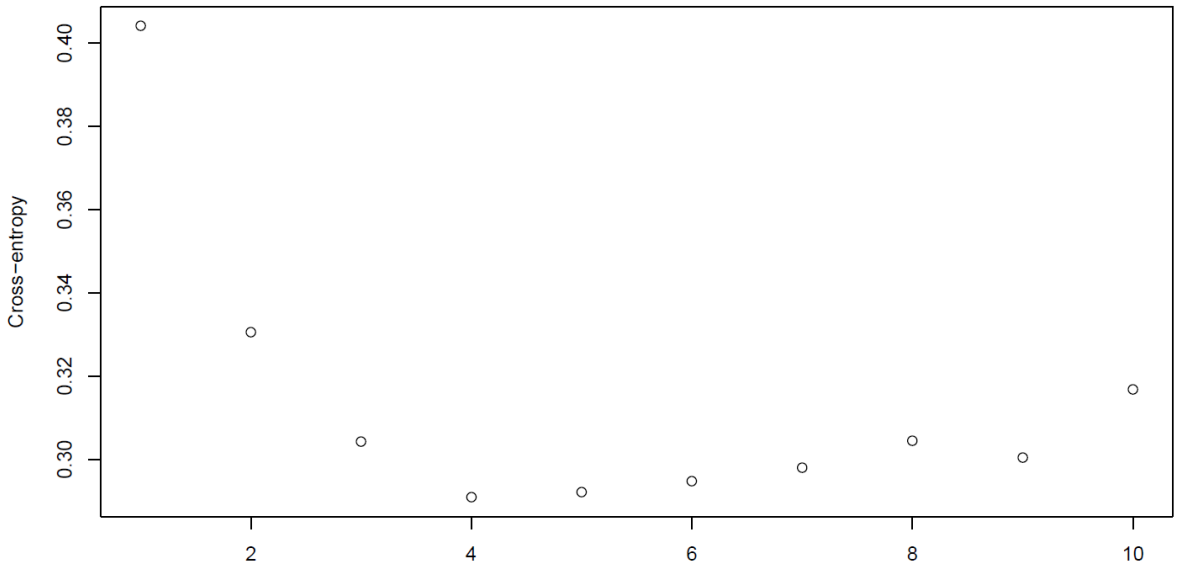
Supplementary Figure S6 (cont.) – Results for the first tree components of the PCA performed on the dataset with only one SNP per block: (A) PC1 and PC2; (B) PC1 and PC3; (C) PC2 and PC3. Each point corresponds to one individual. The clusters formed are the same obtained with the larger dataset (Fig. 3.2). The percentage of the variance explained by the first principal component slightly increases when compared to Fig. 3.2.



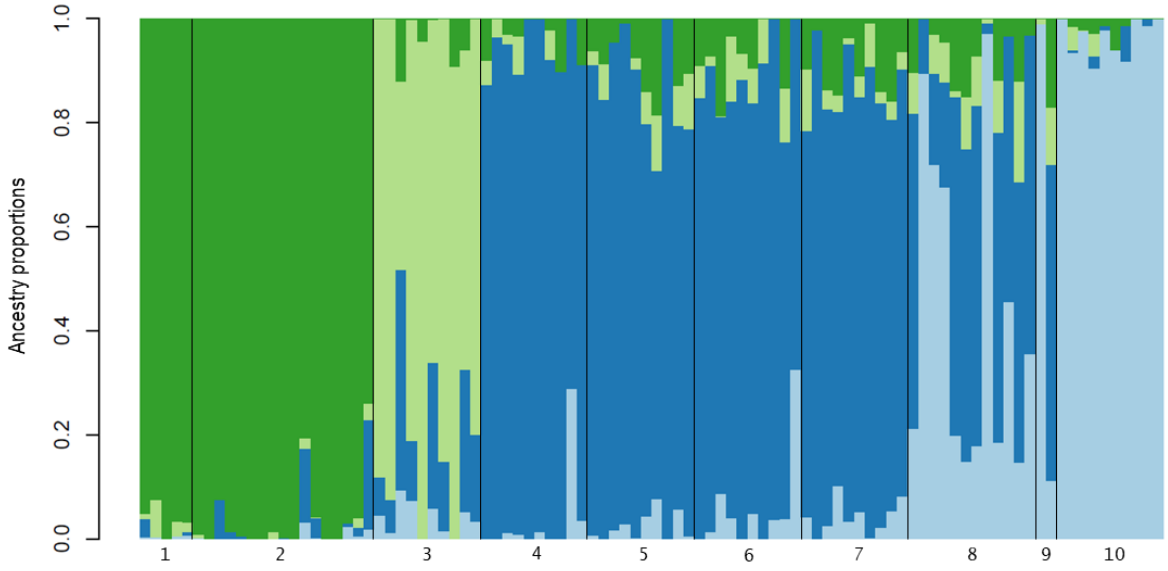
Supplementary Figure S7 – Percentage of variance explained by each principal component (PC) on the Principal Components Analysis performed with a reduced dataset of one SNP per block (Supplementary Figure 6). Together, the first three components explain ≈36% of the variance.



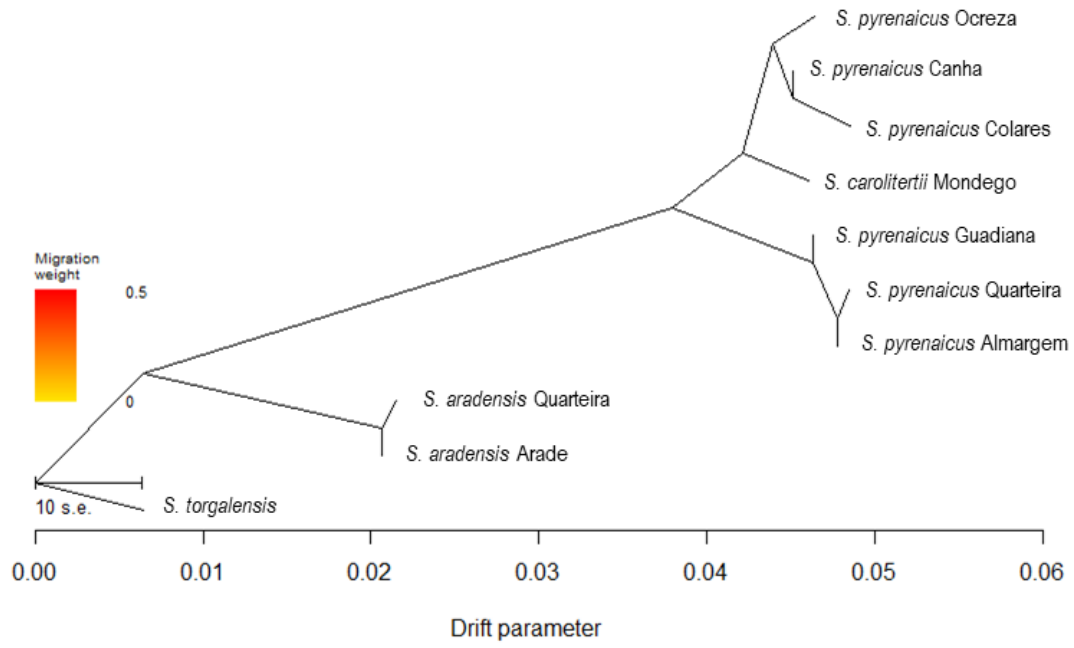
Supplementary Figure S8 – p-values of principal components on the PCA (p<0.01) performed with a reduced dataset of one SNP per block (Supplementary Figure 6). The y-axis shows the p-values of each principal component (x-axis) in log10 scale. Significance of each principal component was assessed with the Tracy-Widom test. The dashed blue line represent the significance level of 0.01 (log10(0.01)=-2). Principal components below the blue line are considered significant.



Supplementary Figure S9 – Cross-entropy for each number of ancestral populations K when sNMF was performed on the dataset with only one SNP per block. The lowest cross-entropy value indicates the number of ancestral populations that best describes the data – in this case, four, as was the case for the dataset with all SNPs.



Supplementary Figure S10 - Ancestry proportions inferred with sNMF for four ancestral populations (K=4) for the dataset with one SNP per block. Each vertical bar corresponds to one individual and the proportion of each colour corresponds to the estimated ancestry proportion from a given cluster. The individuals are grouped per sampling locations and the groups are separated by black lines. Each number corresponds to a sampling location: (1) *S. aradensis* Arade; (2) *S. aradensis* Quarteira; (3) *S. torgalensis*; (4) *S. carolitertii*; (5) *S. pyrenaicus* Ocreza; (6) *S. pyrenaicus* Canha; (7) *S. pyrenaicus* Lizandro; (8) *S. pyrenaicus* Almargem; (9) *S. pyrenaicus* Guadiana; (10) *S. pyrenaicus* Quarteira. The clusters identified are the same as in Fig. 3.3.



Supplementary Figure S11 – Species tree graph obtained with TreeMix for the dataset with one SNP per block. This is an unrooted tree and branch lengths are represented in units of genetic drift, i.e. the longer a given branch the stronger the genetic drift experienced during that branch, which could be due to longer divergence times and/or smaller effective sizes.

Supplementary Table S1 – Detailed sampling locations with GPS coordinates and fishing licences. Colours correspond to those of the species distribution on Figure 1.1.

Sample Code	Species	Basin	Population		Sampling site coordenates	Fishing Licence
CCC_17	*S. carolitertii*	Mondego	Mondego	Sótão river	40°8'5.22"N; 8°8'35.06"W	144/2012/CAPT (Jesus et al. 2016)
CCC_19	*S. carolitetii*			Sótão river	40°8'5.22"N; 8°8'35.06"W	144/2012/CAPT (Jesus et al. 2016)
CCC_11	*S. carolitertii*			Sótão river	40°8'5.22"N; 8°8'35.06"W	144/2012/CAPT (Jesus et al. 2016)
CCC_12	*S. carolitertii*			Sótão river	40°8'5.22"N; 8°8'35.06"W	144/2012/CAPT (Jesus et al. 2016)
CCC_16	*S. carolitertii*			Sótão river	40°8'5.22"N; 8°8'35.06"W	144/2012/CAPT (Jesus et al. 2016)
CCC_2	*S. carolitertii*			Sótão river	40°8'5.22"N; 8°8'35.06"W	144/2012/CAPT (Jesus et al. 2016)
CCC_3	*S. carolitertii*			Sótão river	40°8'5.22"N; 8°8'35.06"W	144/2012/CAPT (Jesus et al. 2016)
CCC_4	*S. carolitertii*			Sótão river	40°8'5.22"N; 8°8'35.06"W	144/2012/CAPT (Jesus et al. 2016)
CCC_5	*S. carolitertii*			Sótão river	40°8'5.22"N; 8°8'35.06"W	144/2012/CAPT (Jesus et al. 2016)
CCC_6	*S. carolitertii*			Sótão river	40°8'5.22"N; 8°8'35.06"W	144/2012/CAPT (Jesus et al. 2016)
PTF1	*S. pyrenaicus*	Tejo	Ocreza	Cobrão stream	39°43'48.22"N; 7°45'38.13" W	142/2012/CAPT (Matos et al. 2015)
PTF2	*S. pyrenaicus*			Cobrão stream	39°43'48.22"N; 7°45'38.13" W	142/2012/CAPT (Matos et al. 2015)
PTF3	*S. pyrenaicus*			Cobrão stream	39°43'48.22"N; 7°45'38.13" W	142/2012/CAPT (Matos et al. 2015)
PTF4	*S. pyrenaicus*			Cobrão stream	39°43'48.22"N; 7°45'38.13" W	142/2012/CAPT (Matos et al. 2015)
PTM1	*S. pyrenaicus*			Cobrão stream	39°43'48.22"N; 7°45'38.13" W	142/2012/CAPT (Matos et al. 2015)
PTM2	*S. pyrenaicus*			Cobrão stream	39°43'48.22"N; 7°45'38.13" W	142/2012/CAPT (Matos et al. 2015)
PTM3	*S. pyrenaicus*			Cobrão stream	39°43'48.22"N; 7°45'38.13" W	142/2012/CAPT (Matos et al. 2015)
PTM4	*S. pyrenaicus*			Cobrão stream	39°43'48.22"N; 7°45'38.13" W	142/2012/CAPT (Matos et al. 2015)
PTM5	*S. pyrenaicus*			Cobrão stream	39°43'48.22"N; 7°45'38.13" W	142/2012/CAPT (Matos et al. 2015)
PTM6	*S. pyrenaicus*			Cobrão stream	39°43'48.22"N; 7°45'38.13" W	142/2012/CAPT (Matos et al. 2015)
SpTs1	*S. pyrenaicus*		Canha	Canha stream	38°44'59.00"N; 8°33'54.00"W	237/2013/CAPT
SpTs2	*S. pyrenaicus*			Canha stream	38°44'59.00"N; 8°33'54.00"W	237/2013/CAPT
SpTs3	*S. pyrenaicus*			Canha stream	38°44'59.00"N; 8°33'54.00"W	237/2013/CAPT
SpTs4	*S. pyrenaicus*			Canha stream	38°44'59.00"N; 8°33'54.00"W	237/2013/CAPT
SpTs5	*S. pyrenaicus*			Canha stream	38°44'59.00"N; 8°33'54.00"W	237/2013/CAPT
SpTs6	*S. pyrenaicus*			Canha stream	38°44'59.00"N; 8°33'54.00"W	237/2013/CAPT
SpTs7	*S. pyrenaicus*			Canha stream	38°44'59.00"N; 8°33'54.00"W	237/2013/CAPT
SpTs8	*S. pyrenaicus*			Canha stream	38°44'59.00"N; 8°33'54.00"W	237/2013/CAPT
SpTs9	*S. pyrenaicus*			Canha stream	38°44'59.00"N; 8°33'54.00"W	237/2013/CAPT
SpTs10	*S. pyrenaicus*			Canha stream	38°44'59.00"N; 8°33'54.00"W	237/2013/CAPT

Supplementary Table S1 (continued)

SpC4	*S. pyrenaicus*	Lizandro	Lizandro	Lizandro river	38°53'18.62''N; 9°19'47.95''W	179/2011/CAPT ; 180/2011/CAPT; 181/2011/CAPT (Inácio et al. 2012)
SpC6	*S. pyrenaicus*			Lizandro river	38°53'18.62''N; 9°19'47.95''W	179/2011/CAPT ; 180/2011/CAPT; 181/2011/CAPT (Inácio et al. 2012)
SpC7	*S. pyrenaicus*			Lizandro river	38°53'18.62''N; 9°19'47.95''W	179/2011/CAPT ; 180/2011/CAPT; 181/2011/CAPT (Inácio et al. 2012)
SpC8	*S. pyrenaicus*			Lizandro river	38°53'18.62''N; 9°19'47.95''W	179/2011/CAPT ; 180/2011/CAPT; 181/2011/CAPT (Inácio et al. 2012)
Pe7	*S. pyrenaicus*			Lizandro river	38°53'18.62''N; 9°19'47.95''W	179/2011/CAPT ; 180/2011/CAPT; 181/2011/CAPT (Inácio et al. 2012)
SpC2	*S. pyrenaicus*			Lizandro river	38°53'18.62''N; 9°19'47.95''W	179/2011/CAPT ; 180/2011/CAPT; 181/2011/CAPT (Inácio et al. 2012)
SpC3	*S. pyrenaicus*			Lizandro river	38°53'18.62''N; 9°19'47.95''W	179/2011/CAPT ; 180/2011/CAPT; 181/2011/CAPT (Inácio et al. 2012)
SpC5	*S. pyrenaicus*			Lizandro river	38°53'18.62''N; 9°19'47.95''W	179/2011/CAPT ; 180/2011/CAPT; 181/2011/CAPT (Inácio et al. 2012)
SpC9	*S. pyrenaicus*			Lizandro river	38°53'18.62''N; 9°19'47.95''W	179/2011/CAPT ; 180/2011/CAPT; 181/2011/CAPT (Inácio et al. 2012)
SpC1	*S. pyrenaicus*			Lizandro river	38°53'18.62''N; 9°19'47.95''W	179/2011/CAPT ; 180/2011/CAPT; 181/2011/CAPT (Inácio et al. 2012)
PSM2	*S. pyrenaicus*	Guadiana	Guadiana	Oeiras stream	37°37'30.29''N; 7°48'37,02''W	235/2013/CAPT (Machado et al. 2016)
PSM3	*S. pyrenaicus*			Oeiras stream	37°37'30.29''N; 7°48'37,02''W	235/2013/CAPT (Machado et al. 2016)
PSF1	*S. pyrenaicus*	Almargem	Almargem	Almargem stream	37°9'50.63''N; 7°37'13.25''W	180/2011/CAPT (Machado et al. 2015)
PSF2	*S. pyrenaicus*			Almargem stream	37°9'50.63''N; 7°37'13.25''W	180/2011/CAPT (Machado et al. 2015)
PSF3	*S. pyrenaicus*			Almargem stream	37°9'50.63''N; 7°37'13.25''W	180/2011/CAPT (Machado et al. 2015)
PSF4	*S. pyrenaicus*			Almargem stream	37°9'50.63''N; 7°37'13.25''W	180/2011/CAPT (Machado et al. 2015)
PSF5	*S. pyrenaicus*			Almargem stream	37°9'50.63''N; 7°37'13.25''W	180/2011/CAPT (Machado et al. 2015)
PSM1	*S. pyrenaicus*			Almargem stream	37°9'50.63''N; 7°37'13.25''W	180/2011/CAPT (Machado et al. 2015)
PSM4	*S. pyrenaicus*			Almargem stream	37°9'50.63''N; 7°37'13.25''W	180/2011/CAPT (Machado et al. 2015)
PSM5	*S. pyrenaicus*			Almargem stream	37°9'50.63''N; 7°37'13.25''W	180/2011/CAPT (Machado et al. 2015)
PP1	*S. pyrenaicus*			Almargem stream	37°9'50.63''N; 7°37'13.25''W	180/2011/CAPT (Machado et al. 2015)
PP2	*S. pyrenaicus*			Almargem stream	37°9'50.63''N; 7°37'13.25''W	180/2011/CAPT (Machado et al. 2015)
PP3	*S. pyrenaicus*			Almargem stream	37°9'50.63''N; 7°37'13.25''W	180/2011/CAPT (Machado et al. 2015)
PP4	*S. pyrenaicus*			Almargem stream	37°9'50.63''N; 7°37'13.25''W	180/2011/CAPT (Machado et al. 2015)

Supplementary Table S1 (continued)

Sample Code	Species	Basin	Population		Sampling site coordenates	Fishing Licence
P38	*S. pyrenaicus*	Quarteira	Quarteira	Quarteira stream	37°13'35.9"; N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)
P66	*S. pyrenaicus*			Quarteira stream	37°13'35.9"; N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)
P72	*S. pyrenaicus*			Quarteira stream	37°13'35.9"; N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)
P74	*S. pyrenaicus*			Quarteira stream	37°13'35.9"; N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)
Q13	*S. pyrenaicus*			Quarteira stream	37°13'35.9"; N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)
Q27	*S. pyrenaicus*			Quarteira stream	37°13'35.9"; N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)
Q49	*S. pyrenaicus*			Quarteira stream	37°13'35.9"; N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)
Q56	*S. pyrenaicus*			Quarteira stream	37°13'35.9"; N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)
Q58	*S. pyrenaicus*			Quarteira stream	37°13'35.9"; N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)
Q59	*S. pyrenaicus*			Quarteira stream	37°13'35.9"; N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)
CCT_1	*S. torgalensis*	Mira	Mira	Torgal stream	37°38'1.31"N; 8°37'22.37"W	144/2012/CAPT (Jesus et al. 2016)
CCT_3	*S. torgalensis*			Torgal stream	37°38'1.31"N; 8°37'22.37"W	144/2012/CAPT (Jesus et al. 2016)
CCT_5	*S. torgalensis*			Torgal stream	37°38'1.31"N; 8°37'22.37"W	144/2012/CAPT (Jesus et al. 2016)
CCT_6	*S. torgalensis*			Torgal stream	37°38'1.31"N; 8°37'22.37"W	144/2012/CAPT (Jesus et al. 2016)
CCT_13	*S. torgalensis*			Torgal stream	37°38'1.31"N; 8°37'22.37"W	144/2012/CAPT (Jesus et al. 2016)
CCT_2	*S. torgalensis*			Torgal stream	37°38'1.31"N; 8°37'22.37"W	144/2012/CAPT (Jesus et al. 2016)
CCT_4	*S. torgalensis*			Torgal stream	37°38'1.31"N; 8°37'22.37"W	144/2012/CAPT (Jesus et al. 2016)
CCT_7	*S. torgalensis*			Torgal stream	37°38'1.31"N; 8°37'22.37"W	144/2012/CAPT (Jesus et al. 2016)
CCT_16	*S. torgalensis*			Torgal stream	37°38'1.31"N; 8°37'22.37"W	144/2012/CAPT (Jesus et al. 2016)
CCT_18	*S. torgalensis*			Torgal stream	37°38'1.31"N; 8°37'22.37"W	144/2012/CAPT (Jesus et al. 2016)
p73	*S. aradensis*	Arade	Arade	Odelouca stream	37°20'37.64"N; 8°29'02.85"W	235/2013/CAPT; 262/2014/CAPT (Matos et al. 2016)
p84	*S. aradensis*			Odelouca stream	37°20'37.64"N; 8°29'02.85"W	235/2013/CAPT; 262/2014/CAPT (Matos et al. 2016)
p81	*S. aradensis*			Odelouca stream	37°20'37.64"N; 8°29'02.85"W	235/2013/CAPT; 262/2014/CAPT (Matos et al. 2016)
p60	*S. aradensis*			Odelouca stream	37°20'37.64"N; 8°29'02.85"W	235/2013/CAPT; 262/2014/CAPT (Matos et al. 2016)
AA1	*S. aradensis*			Odelouca stream	37°20'37.64"N; 8°29'02.85"W	235/2013/CAPT; 262/2014/CAPT (Matos et al. 2016)

Supplementary Table S1 (continued)

Sample Code	Species	Basin	Population		Sampling site coordenates	Fishing Licence
P01	*S. aradensis*	Quarteira	Quarteira	Quarteira stream	37°13'35.9"N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)
P09	*S. aradensis*			Quarteira stream	37°13'35.9"N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)
P10	*S. aradensis*			Quarteira stream	37°13'35.9"N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)
P18	*S. aradensis*			Quarteira stream	37°13'35.9"N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)
P20	*S. aradensis*			Quarteira stream	37°13'35.9"N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)
P22	*S. aradensis*			Quarteira stream	37°13'35.9"N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)
P33	*S. aradensis*			Quarteira stream	37°13'35.9"N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)
P34	*S. aradensis*			Quarteira stream	37°13'35.9"N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)
P39	*S. aradensis*			Quarteira stream	37°13'35.9"N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)
P41	*S. aradensis*			Quarteira stream	37°13'35.9"N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)
P48	*S. aradensis*			Quarteira stream	37°13'35.9"N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)
P53	*S. aradensis*			Quarteira stream	37°13'35.9"N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)
P56	*S. aradensis*			Quarteira stream	37°13'35.9"N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)
Q18	*S. aradensis*			Quarteira stream	37°13'35.9"N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)
Q32	*S. aradensis*			Quarteira stream	37°13'35.9"N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)
Q47	*S. aradensis*			Quarteira stream	37°13'35.9"N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)
Q50	*S. aradensis*			Quarteira stream	37°13'35.9"N 8°01'54.9"W	140/2012/CAPT (Morgado-Santos et al. 2018)

Supplementary Table S2 – Individual median depth of coverage after mapping all reads against the catalogue. Colours correspond to those of the species distribution on Figure 1.1.

Sample Code	Species	Population	Median coverage per individual
CCC_17	*S. carolitertii*	Mondego	86x
CCC_19	*S. carolitertii*		49x
CCC_11	*S. carolitertii*		49x
CCC_12	*S. carolitertii*		66x
CCC_16	*S. carolitertii*		60x
CCC_2	*S. carolitertii*		104x
CCC_3	*S. carolitertii*		60x
CCC_4	*S. carolitertii*		86x
CCC_5	*S. carolitertii*		50x
CCC_6	*S. carolitertii*		44x
PTF1	*S. pyrenaicus*	Ocreza	27x
PTF2	*S. pyrenaicus*		44x
PTF3	*S. pyrenaicus*		51x
PTF4	*S. pyrenaicus*		62x
PTM1	*S. pyrenaicus*		39x
PTM2	*S. pyrenaicus*		32x
PTM3	*S. pyrenaicus*		41x
PTM4	*S. pyrenaicus*		66x
PTM5	*S. pyrenaicus*		34x
PTM6	*S. pyrenaicus*		31x
SpTs1	*S. pyrenaicus*	Canha	41x
SpTs2	*S. pyrenaicus*		42x
SpTs3	*S. pyrenaicus*		34x
SpTs4	*S. pyrenaicus*		49x
SpTs5	*S. pyrenaicus*		47x
SpTs6	*S. pyrenaicus*		132x
SpTs7	*S. pyrenaicus*		42x
SpTs8	*S. pyrenaicus*		87x
SpTs9	*S. pyrenaicus*		58x
SpTs10	*S. pyrenaicus*		49x
SpC4	*S. pyrenaicus*	Lizandro	44x
SpC6	*S. pyrenaicus*		37x
SpC7	*S. pyrenaicus*		32x
SpC8	*S. pyrenaicus*		37x
Pe7	*S. pyrenaicus*		32x
SpC2	*S. pyrenaicus*		37x
SpC3	*S. pyrenaicus*		39x
SpC5	*S. pyrenaicus*		37x
SpC9	*S. pyrenaicus*		38x
SpC1	*S. pyrenaicus*		55x
PSM2	*S. pyrenaicus*	Guadiana	33x
PSM3	*S. pyrenaicus*		132x

Supplementary Table S2 (continued)

Sample Code	Species	Population	Median coverage per individual
PSF1	*S. pyrenaicus*	Almargem	30x
PSF2	*S. pyrenaicus*		100x
PSF3	*S. pyrenaicus*		40x
PSF4	*S. pyrenaicus*		40x
PSF5	*S. pyrenaicus*		45x
PSM1	*S. pyrenaicus*		56x
PSM4	*S. pyrenaicus*		30x
PSM5	*S. pyrenaicus*		106x
PP1	*S. pyrenaicus*		17x
PP2	*S. pyrenaicus*		9x
PP3	*S. pyrenaicus*		10x
PP4	*S. pyrenaicus*		29x
P38	*S. pyrenaicus*	Quarteira	15x
P66	*S. pyrenaicus*		43x
P72	*S. pyrenaicus*		19x
P74	*S. pyrenaicus*		26x
Q13	*S. pyrenaicus*		28x
Q27	*S. pyrenaicus*		21x
Q49	*S. pyrenaicus*		25x
Q56	*S. pyrenaicus*		31x
Q58	*S. pyrenaicus*		31x
Q59	*S. pyrenaicus*		17x
CCT_1	*S. torgalensis*	Mira	60x
CCT_3	*S. torgalensis*		40x
CCT_5	*S. torgalensis*		67x
CCT_6	*S. torgalensis*		48x
CCT_13	*S. torgalensis*		70x
CCT_2	*S. torgalensis*		101x
CCT_4	*S. torgalensis*		76x
CCT_7	*S. torgalensis*		56x
CCT_16	*S. torgalensis*		49x
CCT_18	*S. torgalensis*		50x
p73	*S. aradensis*	Arade	24x
p84	*S. aradensis*		25x
p81	*S. aradensis*		33x
p60	*S. aradensis*		36x
AA1	*S. aradensis*		21x

Supplementary Table S2 (continued)

Sample Code	Species	Population	Median coverage per individual
P01	*S. aradensis*	Quarteira	52x
P09	*S. aradensis*		23x
P10	*S. aradensis*		22x
P18	*S. aradensis*		19x
P20	*S. aradensis*		22x
P22	*S. aradensis*		9x
P33	*S. aradensis*		31x
P34	*S. aradensis*		36x
P39	*S. aradensis*		24x
P41	*S. aradensis*		18x
P48	*S. aradensis*		37x
P53	*S. aradensis*		15x
P56	*S. aradensis*		36x
Q18	*S. aradensis*		9x
Q32	*S. aradensis*		32x
Q47	*S. aradensis*		33x
Q50	*S. aradensis*		27x

Supplementary Table S3 – Number of SNPs per sampling location that significantly deviate from Hardy-Weinberg equilibrium (p<0.05) due to a deficit (A) or excess (B) of heterozygotes for the different filtering options. Colours correspond to those of the species distribution on Figure 1.1.

(A)

Population	No filters	MAF ≥ 0,01	MAF ≥ 0,01 + ⅓ to 2x DP median	MAF ≥ 0,01 + ¼ to 4x DP median
S. carolitertii	23	23	72	72
S. pyrenaicus Ocreza	12	12	138	192
S. pyrenaicus Lizandro	12	12	187	236
S. pyrenaicus Canha	15	15	61	62
S. pyrenaicus Guadiana	0	0	0	0
S. pryrenaicus Almargem	44	44	39	64
S. pyrenaicus Quarteira	10	10	0	1
S. torgalensis	16	16	68	45
S. aradensis Arade	9	9	0	0
S. aradensis Quarteira	17	17	1	5

(B)

Population	No filters	MAF ≥ 0,01	MAF ≥ 0,01 + ⅓ to 2x DP median	MAF ≥ 0,01 + ¼ to 4x DP median
S. carolitertii	61	61	1	6
S. pyrenaicus Ocreza	16	16	0	1
S. pyrenaicus Lizandro	19	19	0	0
S. pyrenaicus Canha	30	30	1	1
S. pyrenaicus Guadiana	0	0	0	0
S. pryrenaicus Almargem	51	51	7	14
S. pyrenaicus Quarteira	1752	1752	653	1057
S. torgalensis	56	56	2	12
S. aradensis Arade	0	0	0	0
S. aradensis Quarteira	2817	2817	1178	1904

Supplementary Table S4 – Percentage of missing data of each individual on the final dataset. Colours correspond to those of the species distribution on Figure 1.1.

Sample Code	Species	Population	Percentage of missing data (MAF≥0,01 + 1/4 to 4x DP median)
CCC_17	*S. carolitertii*	Mondego	31.88
CCC_19	*S. carolitertii*		48.86
CCC_11	*S. carolitertii*		44.98
CCC_12	*S. carolitertii*		35.76
CCC_16	*S. carolitertii*		38.86
CCC_2	*S. carolitertii*		37.88
CCC_3	*S. carolitertii*		35.62
CCC_4	*S. carolitertii*		30.42
CCC_5	*S. carolitertii*		43.41
CCC_6	*S. carolitertii*		48.57
PTF1	*S. pyrenaicus*	Ocreza	59.19
PTF2	*S. pyrenaicus*		41.54
PTF3	*S. pyrenaicus*		38.99
PTF4	*S. pyrenaicus*		35.91
PTM1	*S. pyrenaicus*		48.19
PTM2	*S. pyrenaicus*		56.07
PTM3	*S. pyrenaicus*		45.67
PTM4	*S. pyrenaicus*		32.43
PTM5	*S. pyrenaicus*		52.19
PTM6	*S. pyrenaicus*		69.85
SpTs1	*S. pyrenaicus*	Canha	50.49
SpTs2	*S. pyrenaicus*		46.46
SpTs3	*S. pyrenaicus*		62.54
SpTs4	*S. pyrenaicus*		44.07
SpTs5	*S. pyrenaicus*		51.41
SpTs6	*S. pyrenaicus*		26.85
SpTs7	*S. pyrenaicus*		53.66
SpTs8	*S. pyrenaicus*		36.19
SpTs9	*S. pyrenaicus*		39.22
SpTs10	*S. pyrenaicus*		43.29
SpC4	*S. pyrenaicus*	Lizandro	45.62
SpC6	*S. pyrenaicus*		52.25
SpC7	*S. pyrenaicus*		59.13
SpC8	*S. pyrenaicus*		57.78
Pe7	*S. pyrenaicus*		53.92
SpC2	*S. pyrenaicus*		49.86
SpC3	*S. pyrenaicus*		41.78
SpC5	*S. pyrenaicus*		52.53
SpC9	*S. pyrenaicus*		54.70
SpC1	*S. pyrenaicus*		36.33
PSM2	*S. pyrenaicus*	Guadiana	69.27
PSM3	*S. pyrenaicus*		40.46

Supplementary Table S4 (continued)

Sample Code	Species	Population	Percentage of missing data (MAF≥0,01 + 1/4 to 4x DP median)
PSF1	*S. pyrenaicus*	Almargem	67.88
PSF2	*S. pyrenaicus*		30.98
PSF3	*S. pyrenaicus*		49.71
PSF4	*S. pyrenaicus*		46.82
PSF5	*S. pyrenaicus*		46.88
PSM1	*S. pyrenaicus*		45.40
PSM4	*S. pyrenaicus*		65.00
PSM5	*S. pyrenaicus*		30.30
PP1	*S. pyrenaicus*		36.31
PP2	*S. pyrenaicus*		45.04
PP3	*S. pyrenaicus*		43.84
PP4	*S. pyrenaicus*		34.75
P38	*S. pyrenaicus*	Quarteira	33.80
P66	*S. pyrenaicus*		31.53
P72	*S. pyrenaicus*		33.16
P74	*S. pyrenaicus*		33.49
Q13	*S. pyrenaicus*		33.03
Q27	*S. pyrenaicus*		34.71
Q49	*S. pyrenaicus*		34.79
Q56	*S. pyrenaicus*		32.94
Q58	*S. pyrenaicus*		34.45
Q59	*S. pyrenaicus*		37.15
CCT_1	*S. torgalensis*	Mira	41.56
CCT_3	*S. torgalensis*		60.81
CCT_5	*S. torgalensis*		40.59
CCT_6	*S. torgalensis*		48.36
CCT_13	*S. torgalensis*		37.05
CCT_2	*S. torgalensis*		40.78
CCT_4	*S. torgalensis*		36.51
CCT_7	*S. torgalensis*		39.63
CCT_16	*S. torgalensis*		49.35
CCT_18	*S. torgalensis*		45.85
p73	*S. aradensis*	Arade	35.20
p84	*S. aradensis*		35.99
p81	*S. aradensis*		34.98
p60	*S. aradensis*		34.13
AA1	*S. aradensis*		37.49

Supplementary Table S4 (continued)

Sample Code	Species	Population	Percentage of missing data (MAF≥0,01 + 1/4 to 4x DP median)
P01	*S. aradensis*	Quarteira	34.49
P09	*S. aradensis*		35.79
P10	*S. aradensis*		35.29
P18	*S. aradensis*		34.99
P20	*S. aradensis*		34.78
P22	*S. aradensis*		46.22
P33	*S. aradensis*		35.34
P34	*S. aradensis*		33.38
P39	*S. aradensis*		34.38
P41	*S. aradensis*		38.81
P48	*S. aradensis*		34.61
P53	*S. aradensis*		36.59
P56	*S. aradensis*		33.52
Q18	*S. aradensis*		47.76
Q32	*S. aradensis*		34.63
Q47	*S. aradensis*		37.76
Q50	*S. aradensis*		36.49

Supplementary Table S5 – Number of polymorphic and monomorphic sites, missing data, private sites and fixed differences within each sampling locations. The values are given in number of sites. Colours correspond to those of the species distribution on Figure 1.1.

	Polymorphic	Monomorphic	Without data	Single individual	Private	Fixed differences
S. carolitertii	7678	19401	624	110	1103	356
***S. pyrenaicus* Ocreza**	6651	20493	559	105	811	245
***S. pyrenaicus* Lizandro**	5745	21287	671	102	447	317
***S. pyrenaicus* Canha**	7475	19855	373	133	670	238
***S. pyrenaicus* Guadiana**	2895	17166	7642	2397	27	1100
***S. pyrenaicus* Almargem**	8063	19210	430	78	983	285
***S. pyrenaicus* Quarteira**	7833	15603	4267	190	815	908
S. torgalensis	6646	20218	839	69	2020	540
***S. aradensis* Arade**	5489	16149	6065	368	143	1232
***S. aradensis* Quarteira**	7825	15793	4085	254	1073	978

Supplementary Table S6 – Mean expected heterozygosity and mean observed heterozygosity across sites for each sampling location. Colours correspond to those of the species distribution on Figure 1.1.

Population	Mean expected Heterozygosity	Mean observed Heterozygosity
S. carolitertii	0.079	0.074
S. pyrenaicus Ocreza	0.070	0.047
S. pyrenaicus Lizandro	0.062	0.038
S. pyrenaicus Canha	0.079	0.066
S. pyrenaicus Guadiana	0.093	0.083
S. pyrenaicus Almargem	0.087	0.086
S. pyrenaicus Quarteira	0.106	0.154
S. torgalensis	0.083	0.085
S. aradensis Arade	0.107	0.156
S. aradensis Quarteira	0.102	0.153

Supplementary Table S7 – Quantiles 5% and 95% for the distribution of the expected and observed heterozygosity across sites for each sampling location. Colours correspond to those of the species distribution on Figure 1.1.

	Mean Expected Heterozygosity		Mean Observed Heterozygosity	
	Quantile 5%	Quantile 95%	Quantile 5%	Quantile 95%
S. carolitertii	0.031	0.152	0.046	0.122
S. pyrenaicus Ocreza	0.031	0.136	0.026	0.076
S. pyrenaicus Lizandro	0.027	0.121	0.027	0.058
S. pyrenaicus Canha	0.034	0.151	0.031	0.139
S. pyrenaicus Guadiana	0.072	0.144	0.034	0.131
S. pyrenaicus Almargem	0.033	0.158	0.024	0.146
S. pyrenaicus Quarteira	0.033	0.182	0.150	0.160
S. torgalensis	0.031	0.135	0.046	0.155
S. aradensis Arade	0.051	0.169	0.154	0.158
S. aradensis Quarteira	0.019	0.184	0.141	0.159

Supplementary Table S8 – Number of polymorphic and monomorphic sites, missing data, private sites and fixed differences within each group. The values are given in number of sites. The order of the groups corresponds to the one on Figure 3.4 (left to right).

	Polymorphic	Monomorphic	Without data	Single individual	Private	Fixed differences
S. aradensis	8285	15610	3808	222	2928	1015
S. torgalensis	6646	20218	839	69	2022	540
***S. carolitertii* + *S. pyrenaicus* Ocreza + *S. pyrenaicus* Lizandro + *S. pyrenaicus* Canha**	13134	14486	83	6	6027	169
***S. pyrenaicus* Almargem + *S. pyrenaicus* Guadiana + *S. pyrenaicus* Quarteira**	10954	16474	275	35	3673	250

Supplementary Table S9 – Mean expected heterozygosity and mean observed heterozygosity across sites for each group identified. The order of the groups corresponds to the one on Figure 3.4 (left to right).

Population	Mean expected Heterozygosity	Mean observed Heterozigocity
S. aradensis	0.111	0.153
S. torgalensis	0.090	0.085
S. carolitertii + *S. pyrenaicus* Canha + *S. pyrenaicus* Lizandro + *S. pyrenaicus* Ocreza	0.093	0.059
S. pyrenaicus Almargem + *S. pyrenaicus* Guadiana + *S. pyrenaicus* Quarteira	0.104	0.113

Supplementary Table S10 – Quantiles 5% and 95% for the distribution of the expected heterozygosity and observed heterozygosity across sites for each group identified. The order of the groups corresponds to the one on Figure 3.4 (left to right).

	Mean Expected Heterozygosity		Mean Observed Heterozygosity	
	Quantile 5%	Quantile 95%	Quantile 5%	Quantile 95%
S. aradensis	0.017	0.185	0.142	0.159
S. torgalensis	0.031	0.135	0.046	0.155
***S. carolitertii* + *S. pyrenaicus* Ocreza + *S. pyrenaicus* Lizandro + *S. pyrenaicus* Canha**	0.017	0.235	0.028	0.125
***S. pyrenaicus* Almargem + *S. pyrenaicus* Guadiana + *S. pyrenaicus* Quarteira**	0.021	0.205	0.024	0.153

Supplementary Table S11 – Detailed results of the D-statistic calculated for the different scenarios in Fig. 2.1. Values of D significantly different from zero (p<0.01) are highlighted in grey shading on the p-value column. Possible combinations are grouped per outgroup (P outgroup). For each combination of populations, we report the D statistic values, the corresponding number of ABBA and BABA sites, as well as the estimated standard deviation, z-score and p-value obtained with the block jackknife approach. Colours correspond to the species distribution on Figure 1.1.

Tree A

P1	P2	P3	P outgroup	D	ABBA	BABA	sd	z	p-value
S. pyrenaicus Almargem	*S. pyrenaicus* Ocreza	*S. carolitertii*	*S. torgalensis*	0.209	361.164	236.513	0.016	13.281	0.000
	S. pyrenaicus Lizandro			0.217	363.257	233.720	0.020	10.971	0.000
	S. pyrenaicus Canha			0.220	378.989	242.429	0.013	16.709	0.000

P1	P2	P3	P outgroup	D	ABBA	BABA	sd	z	p-value
S. pyrenaicus Quarteira	*S. pyrenaicus* Ocreza	*S. carolitertii*	*S. torgalensis*	0.218	352.135	226.053	0.015	14.480	0.000
	S. pyrenaicus Lizandro			0.234	356.790	221.701	0.019	12.560	0.000
	S. pyrenaicus Canha			0.237	366.623	226.245	0.014	17.315	0.000

P1	P2	P3	P outgroup	D	ABBA	BABA	sd	z	p-value
S. pyrenaicus Almargem	*S. pyrenaicus* Ocreza	*S. carolitertii*	*S. aradensis* Arade	0.202	305.600	203.030	0.013	16.015	0.000
	S. pyrenaicus Lizandro			0.198	303.327	202.943	0.019	10.257	0.000
	S. pyrenaicus Canha			0.217	322.386	207.228	0.012	18.176	0.000

P1	P2	P3	P outgroup	D	ABBA	BABA	sd	z	p-value
S. pyrenaicus Quarteira	*S. pyrenaicus* Ocreza	*S. carolitertii*	*S. aradensis* Arade	0.250	338.522	203.315	0.014	17.520	0.000
	S. pyrenaicus Lizandro			0.255	337.538	200.372	0.017	15.347	0.000
	S. pyrenaicus Canha			0.271	351.102	201.570	0.014	19.736	0.000

P1	P2	P3	P outgroup	D	ABBA	BABA	sd	z	p-value
S. pyrenaicus Almargem	*S. pyrenaicus* Ocreza	*S. carolitertii*	*S. aradensis* Quarteira	0.205	332.429	219.138	0.015	13.425	0.000
	S. pyrenaicus Lizandro			0.207	333.835	219.113	0.019	10.982	0.000
	S. pyrenaicus Canha			0.219	351.269	225.162	0.014	16.086	0.000

P1	P2	P3	P outgroup	D	ABBA	BABA	sd	z	p-value
S. pyrenaicus Quarteira	*S. pyrenaicus* Ocreza	*S. carolitertii*	*S. aradensis* Quarteira	0.261	366.160	214.603	0.014	18.189	0.000
	S. pyrenaicus Lizandro			0.271	369.770	212.171	0.017	16.270	0.000
	S. pyrenaicus Canha			0.281	381.166	214.143	0.013	21.691	0.000

Supplementary Table S11 (continued)

Tree B

P1	P2	P3	P outgroup	D	ABBA	BABA	sd	z	p-value
S. carolitertii	*S. pyrenaicus* Ocreza	*S. pyrenaicus* Almargem	*S. torgalensis*	0.017	244.586	236.513	0.014	1.161	0.123
	S. pyrenaicus Lizandro			0.028	247.052	233.720	0.022	1.288	0.099
	S. pyrenaicus Canha			0.041	263.035	242.429	0.014	2.894	0.002

P1	P2	P3	P outgroup	D	ABBA	BABA	sd	z	p-value
S. carolitertii	*S. pyrenaicus* Ocreza	*S. pyrenaicus* Quarteira	*S. torgalensis*	-0.023	215.782	226.053	0.015	-1.535	0.938
	S. pyrenaicus Lizandro			0.000	221.783	221.701	0.019	0.010	0.496
	S. pyrenaicus Canha			0.038	244.106	226.245	0.012	3.300	0.000

P1	P2	P3	P outgroup	D	ABBA	BABA	sd	z	p-value
S. carolitertii	*S. pyrenaicus* Ocreza	*S. pyrenaicus* Almargem	*S. aradensis* Arade	0.008	206.445	203.030	0.016	0.506	0.306
	S. pyrenaicus Lizandro			0.011	207.584	202.943	0.024	0.478	0.316
	S. pyrenaicus Canha			0.039	224.169	207.228	0.018	2.126	0.017

P1	P2	P3	P outgroup	D	ABBA	BABA	sd	z	p-value
S. carolitertii	*S. pyrenaicus* Ocreza	*S. pyrenaicus* Quarteira	*S. aradensis* Arade	-0.027	192.484	203.315	0.017	-1.626	0.948
	S. pyrenaicus Lizandro			-0.012	195.753	200.372	0.023	-0.506	0.693
	S. pyrenaicus Canha			0.039	218.009	201.570	0.014	2.723	0.003

P1	P2	P3	P outgroup	D	ABBA	BABA	sd	z	p-value
S. carolitertii	*S. pyrenaicus* Ocreza	*S. pyrenaicus* Almargem	*S. aradensis* Quarteira	0.016	226.301	219.138	0.016	1.000	0.159
	S. pyrenaicus Lizandro			0.018	227.025	219.113	0.023	0.761	0.223
	S. pyrenaicus Canha			0.038	243.029	225.162	0.018	2.111	0.017

P1	P2	P3	P outgroup	D	ABBA	BABA	sd	z	p-value
S. carolitertii	*S. pyrenaicus* Ocreza	*S. pyrenaicus* Quarteira	*S. aradensis* Quarteira	-0.025	204.083	214.603	0.016	-1.609	0.946
	S. pyrenaicus Lizandro			-0.012	207.338	212.171	0.023	-0.505	0.693
	S. pyrenaicus Canha			0.036	230.070	214.143	0.013	2.780	0.003

Supplementary Table S11 (continued)

Tree C

P1	P2	P3	P outgroup	D	ABBA	BABA	sd	z	p-value
S. carolitertii	*S. pyrenaicus* Ocreza	*S. pyrenaicus* Lizandro	*S. torgalensis*	0.090	318.787	266.130	0.014	6.211	0.000
		S. pyrenaicus Canha		0.091	338.463	282.168	0.015	6.174	0.000

P1	P2	P3	P outgroup	D	ABBA	BABA	sd	z	p-value
S. carolitertii	*S. pyrenaicus* Ocreza	*S. pyrenaicus* Lizandro	*S. aradensis* Arade	0.092	264.548	220.138	0.016	5.730	0.000
		S. pyrenaicus Canha		0.089	283.896	237.656	0.016	5.648	0.000

P1	P2	P3	P outgroup	D	ABBA	BABA	sd	z	p-value
S. carolitertii	*S. pyrenaicus* Ocreza	*S. pyrenaicus* Lizandro	*S. aradensis* Quarteira	0.084	289.149	244.470	0.014	5.832	0.000
		S. pyrenaicus Canha		0.085	308.714	260.395	0.014	6.103	0.000

Tree D

P1	P2	P3	P outgroup	D	ABBA	BABA	sd	z	p-value
S. pyrenaicus Lizandro	*S. pyrenaicus* Ocreza	*S. carolitertii*	*S. torgalensis*	-0.006	263.179	266.130	0.016	-0.357	0.640
S. pyrenaicus Canha				-0.014	274.532	282.168	0.013	-1.038	0.850

P1	P2	P3	P outgroup	D	ABBA	BABA	sd	z	p-value
S. pyrenaicus Lizandro	*S. pyrenaicus* Ocreza	*S. carolitertii*	*S. aradensis* Arade	0.004	221.906	220.138	0.019	0.212	0.416
S. pyrenaicus Canha				-0.018	229.204	237.656	0.010	-1.790	0.963

P1	P2	P3	P outgroup	D	ABBA	BABA	sd	z	p-value
S. pyrenaicus Lizandro	*S. pyrenaicus* Ocreza	*S. carolitertii*	*S. aradensis* Quarteira	-0.001	243.986	244.470	0.018	-0.055	0.522
S. pyrenaicus Canha				-0.017	251.790	260.395	0.010	-1.643	0.950

Supplementary Table S12 - Detailed results of the D-statistic calculated per individual for the different scenarios in Fig. 2.1. Values of D significantly different from zero ($p<0.01$) are highlighted in grey shading on the p-value column. Possible combinations are grouped per outgroup (P outgroup), in the same order as in Supplementary Table S11. For each combination of populations, we report the D-statistic values, as well as the estimated standard deviation, z-score and p-value obtained with the block jackknife approach. Colours correspond to those of the species distribution on Figure 1.1.

P1	P2	P3	P outgroup	D	sd	z	p-value
S. pyrenaicus Almargem	*S. pyrenaicus* Ocreza	*S. carolitertii*	*S. torgalensis*	0.190	0.022	8.491	0.000
				0.185	0.025	7.442	0.000
				0.190	0.019	9.881	0.000
				0.215	0.020	10.698	0.000
				0.174	0.026	6.806	0.000
				0.047	0.045	1.042	0.149
				0.201	0.042	4.788	0.000
				0.222	0.021	10.487	0.000
				0.194	0.035	5.483	0.000
				0.178	0.032	5.553	0.000
	S. pyrenaicus Lizandro			0.201	0.037	5.454	0.000
				0.191	0.018	10.599	0.000
				0.235	0.032	7.468	0.000
				0.131	0.033	3.906	0.000
				0.243	0.026	9.173	0.000
				0.193	0.032	6.124	0.000
				0.176	0.029	6.035	0.000
				0.209	0.027	7.708	0.000
				0.191	0.032	6.041	0.000
				0.188	0.029	6.479	0.000
	S. pyrenaicus Canha			0.170	0.029	5.758	0.000
				0.210	0.023	9.280	0.000
				0.142	0.024	5.854	0.000
				0.120	0.030	3.926	0.000
				0.175	0.022	8.082	0.000
				0.156	0.031	5.113	0.000
				0.211	0.022	9.813	0.000
				0.220	0.015	14.287	0.000
				0.224	0.038	5.893	0.000
				0.215	0.018	12.013	0.000

Supplementary Table S12 (continued)

P1	P2	P3	P outgroup	D	sd	z	p-value
S. pyrenaicus Quarteira	*S. pyrenaicus* Ocreza	*S. carolitertii*	*S. torgalensis*	0.200	0.028	7.098	0.000
				0.183	0.025	7.210	0.000
				0.206	0.022	9.184	0.000
				0.221	0.018	12.003	0.000
				0.212	0.025	8.474	0.000
				0.078	0.041	1.901	0.029
				0.220	0.039	5.590	0.000
				0.226	0.023	9.945	0.000
				0.192	0.036	5.334	0.000
				0.190	0.031	6.108	0.000
	S. pyrenaicus Lizandro			0.229	0.034	6.786	0.000
				0.213	0.021	10.255	0.000
				0.237	0.029	8.213	0.000
				0.167	0.033	5.054	0.000
				0.257	0.027	9.566	0.000
				0.217	0.038	5.678	0.000
				0.207	0.030	6.969	0.000
				0.218	0.028	7.717	0.000
				0.210	0.032	6.468	0.000
				0.202	0.029	6.866	0.000
	S. pyrenaicus Canha			0.183	0.032	5.640	0.000
				0.216	0.023	9.225	0.000
				0.166	0.028	5.925	0.000
				0.157	0.028	5.580	0.000
				0.181	0.023	7.851	0.000
				0.170	0.031	5.551	0.000
				0.223	0.021	10.435	0.000
				0.228	0.013	17.011	0.000
				0.236	0.034	7.025	0.000
				0.238	0.018	13.186	0.000

Supplementary Table S12 (continued)

P1	P2	P3	P outgroup	D	sd	z	p-value
S. pyrenaicus Almargem	*S. pyrenaicus* Ocreza	*S. carolitertii*	*S. aradensis* Arade	0.186	0.028	6.687	0.000
				0.171	0.030	5.763	0.000
				0.194	0.023	8.516	0.000
				0.209	0.024	8.784	0.000
				0.165	0.031	5.327	0.000
				-0.001	0.044	-0.017	0.507
				0.161	0.045	3.571	0.000
				0.223	0.027	8.172	0.000
				0.216	0.036	6.078	0.000
				0.189	0.039	4.847	0.000
	S. pyrenaicus Lizandro			0.169	0.039	4.315	0.000
				0.182	0.024	7.543	0.000
				0.241	0.029	8.364	0.000
				0.148	0.032	4.611	0.000
				0.205	0.027	7.653	0.000
				0.178	0.031	5.811	0.000
				0.186	0.029	6.312	0.000
				0.121	0.035	3.432	0.000
				0.201	0.038	5.327	0.000
				0.190	0.032	5.956	0.000
	S. pyrenaicus Canha			0.186	0.036	5.230	0.000
				0.186	0.030	6.200	0.000
				0.173	0.028	6.211	0.000
				0.132	0.032	4.168	0.000
				0.170	0.026	6.583	0.000
				0.129	0.030	4.270	0.000
				0.220	0.022	9.877	0.000
				0.219	0.017	13.181	0.000
				0.267	0.037	7.227	0.000
				0.179	0.020	8.777	0.000

Supplementary Table S12 (continued)

P1	P2	P3	P outgroup	D	sd	z	p-value
S. pyrenaicus Quarteira	*S. pyrenaicus* Ocreza	*S. carolitertii*	*S. aradensis* Arade	0.225	0.029	7.627	0.000
				0.221	0.029	7.534	0.000
				0.255	0.024	10.477	0.000
				0.268	0.020	13.313	0.000
				0.225	0.028	8.059	0.000
				0.082	0.040	2.046	0.020
				0.211	0.042	4.983	0.000
				0.263	0.026	10.160	0.000
				0.252	0.034	7.451	0.000
				0.239	0.035	6.737	0.000
	S. pyrenaicus Lizandro			0.229	0.037	6.174	0.000
				0.241	0.027	8.927	0.000
				0.283	0.024	11.779	0.000
				0.218	0.029	7.418	0.000
				0.262	0.022	11.813	0.000
				0.232	0.032	7.170	0.000
				0.241	0.027	8.971	0.000
				0.202	0.032	6.244	0.000
				0.267	0.035	7.680	0.000
				0.244	0.029	8.385	0.000
	S. pyrenaicus Canha			0.228	0.035	6.427	0.000
				0.237	0.025	9.603	0.000
				0.221	0.030	7.348	0.000
				0.198	0.030	6.600	0.000
				0.212	0.023	9.093	0.000
				0.183	0.030	6.119	0.000
				0.266	0.022	12.028	0.000
				0.266	0.015	17.660	0.000
				0.312	0.034	9.095	0.000
				0.226	0.019	12.031	0.000

Supplementary Table S12 (continued)

P1	P2	P3	P outgroup	D	sd	z	p-value
S. pyrenaicus Almargem	*S. pyrenaicus* Ocreza	*S. carolitertii*	*S. aradensis* Quarteira	0.186	0.028	6.539	0.000
				0.184	0.026	6.979	0.000
				0.199	0.025	8.113	0.000
				0.209	0.023	9.083	0.000
				0.151	0.032	4.677	0.000
				0.023	0.045	0.516	0.303
				0.152	0.041	3.708	0.000
				0.221	0.024	9.143	0.000
				0.225	0.036	6.193	0.000
				0.200	0.033	6.022	0.000
	S. pyrenaicus Lizandro			0.174	0.041	4.207	0.000
				0.192	0.022	8.778	0.000
				0.229	0.026	8.690	0.000
				0.157	0.031	5.138	0.000
				0.207	0.025	8.302	0.000
				0.201	0.028	7.065	0.000
				0.185	0.032	5.745	0.000
				0.165	0.030	5.472	0.000
				0.203	0.038	5.283	0.000
				0.194	0.028	6.872	0.000
	S. pyrenaicus Canha			0.189	0.032	5.854	0.000
				0.194	0.027	7.219	0.000
				0.171	0.028	6.139	0.000
				0.137	0.030	4.591	0.000
				0.177	0.025	7.070	0.000
				0.141	0.028	5.042	0.000
				0.216	0.021	10.514	0.000
				0.224	0.015	14.508	0.000
				0.282	0.037	7.608	0.000
				0.171	0.019	9.024	0.000

Supplementary Table S12 (continued)

P1	P2	P3	P outgroup	D	sd	z	p-value
S. pyrenaicus Quarteira	*S. pyrenaicus* Ocreza	*S. carolitertii*	*S. aradensis* Quarteira	0.244	0.029	8.485	0.000
				0.237	0.027	8.854	0.000
				0.263	0.024	10.976	0.000
				0.269	0.019	14.068	0.000
				0.229	0.028	8.148	0.000
				0.112	0.040	2.803	0.003
				0.213	0.037	5.817	0.000
				0.271	0.024	11.104	0.000
				0.249	0.034	7.400	0.000
				0.248	0.030	8.341	0.000
	S. pyrenaicus Lizandro			0.244	0.037	6.570	0.000
				0.250	0.024	10.327	0.000
				0.288	0.024	12.230	0.000
				0.226	0.029	7.889	0.000
				0.265	0.023	11.707	0.000
				0.264	0.031	8.575	0.000
				0.260	0.029	8.931	0.000
				0.239	0.028	8.428	0.000
				0.281	0.033	8.424	0.000
				0.260	0.026	9.958	0.000
	S. pyrenaicus Canha			0.236	0.031	7.490	0.000
				0.254	0.023	11.153	0.000
				0.233	0.027	8.555	0.000
				0.228	0.027	8.548	0.000
				0.229	0.022	10.558	0.000
				0.200	0.027	7.342	0.000
				0.279	0.020	14.156	0.000
				0.281	0.015	19.321	0.000
				0.326	0.034	9.609	0.000
				0.224	0.018	12.254	0.000

Supplementary Table S12 (continued)

Tree B

P1	P2	P3	P outgroup	D	sd	z	p-value
S. carolitertii	*S. pyrenaicus* Ocreza	*S. pyrenaicus* Almargem	*S. torgalensis*	0.048	0.027	1.759	0.039
				0.011	0.031	0.349	0.364
				0.022	0.025	0.888	0.187
				0.032	0.024	1.355	0.088
				-0.012	0.029	-0.405	0.657
				-0.073	0.040	-1.827	0.966
				0.014	0.039	0.344	0.366
				0.058	0.024	2.409	0.008
				0.014	0.034	0.425	0.335
				-0.006	0.033	-0.188	0.574
	S. pyrenaicus Lizandro			0.023	0.034	0.670	0.251
				0.042	0.023	1.844	0.033
				0.065	0.032	2.061	0.020
				-0.015	0.032	-0.457	0.676
				0.069	0.031	2.222	0.013
				0.014	0.034	0.403	0.343
				-0.012	0.031	-0.395	0.654
				0.063	0.038	1.672	0.047
				0.019	0.038	0.492	0.311
				0.027	0.026	1.072	0.142
	S. pyrenaicus Canha			0.018	0.030	0.583	0.280
				0.031	0.022	1.383	0.083
				-0.031	0.035	-0.892	0.814
				-0.015	0.034	-0.444	0.672
				0.042	0.029	1.441	0.075
				0.020	0.029	0.684	0.247
				0.034	0.028	1.205	0.114
				0.060	0.019	3.131	0.001
				0.015	0.041	0.355	0.361
				0.085	0.020	4.298	0.000

Supplementary Table S12 (continued)

P1	P2	P3	P outgroup	D	sd	z	p-value
S. carolitertii	*S. pyrenaicus* Ocreza	*S. pyrenaicus* Quarteira	*S. torgalensis*	-0.008	0.026	-0.305	0.620
				-0.034	0.032	-1.054	0.854
				-0.031	0.028	-1.114	0.867
				0.014	0.022	0.663	0.254
				-0.037	0.027	-1.378	0.916
				-0.086	0.043	-1.967	0.975
				-0.038	0.044	-0.855	0.804
				0.020	0.022	0.903	0.183
				-0.027	0.042	-0.651	0.742
				-0.065	0.034	-1.906	0.972
	S. pyrenaicus Lizandro			-0.015	0.035	-0.433	0.667
				0.020	0.022	0.911	0.181
				0.016	0.031	0.515	0.303
				-0.027	0.033	-0.816	0.793
				0.052	0.034	1.509	0.066
				-0.018	0.037	-0.495	0.690
				-0.029	0.032	-0.932	0.824
				0.000	0.039	-0.003	0.501
				-0.010	0.037	-0.260	0.602
				0.032	0.031	1.024	0.153
	S. pyrenaicus Canha			-0.013	0.029	-0.443	0.671
				-0.010	0.021	-0.506	0.694
				-0.032	0.032	-1.002	0.842
				-0.025	0.033	-0.770	0.779
				-0.003	0.029	-0.119	0.547
				0.001	0.030	0.027	0.489
				0.030	0.029	1.024	0.153
				0.033	0.018	1.795	0.036
				0.000	0.052	0.007	0.497
				0.188	0.016	11.717	0.000

Supplementary Table S12 (continued)

P1	P2	P3	P outgroup	D	sd	z	p-value
S. carolitertii	*S. pyrenaicus* Ocreza	*S. pyrenaicus* Almargem	*S. aradensis* Arade	0.044	0.032	1.386	0.083
				-0.004	0.045	-0.100	0.540
				0.010	0.026	0.379	0.352
				0.042	0.025	1.721	0.043
				-0.019	0.034	-0.568	0.715
				-0.085	0.048	-1.778	0.962
				0.003	0.044	0.076	0.470
				0.060	0.029	2.050	0.020
				-0.009	0.037	-0.254	0.600
				-0.024	0.040	-0.596	0.724
	S. pyrenaicus Lizandro			-0.022	0.040	-0.550	0.709
				0.019	0.029	0.668	0.252
				0.105	0.035	3.004	0.001
				-0.022	0.037	-0.607	0.728
				0.043	0.030	1.411	0.079
				-0.006	0.043	-0.152	0.560
				-0.002	0.035	-0.044	0.518
				0.003	0.050	0.070	0.472
				0.007	0.045	0.153	0.439
				0.023	0.026	0.889	0.187
	S. pyrenaicus Canha			0.014	0.032	0.450	0.326
				0.025	0.025	0.981	0.163
				0.017	0.038	0.457	0.324
				-0.004	0.034	-0.132	0.552
				0.036	0.033	1.081	0.140
				0.015	0.031	0.486	0.314
				0.059	0.029	2.058	0.020
				0.074	0.020	3.736	0.000
				0.058	0.044	1.309	0.095
				0.032	0.019	1.696	0.045

Supplementary Table S12 (continued)

P1	P2	P3	P outgroup	D	sd	z	p-value
S. carolitertii	*S. pyrenaicus* Ocreza	*S. pyrenaicus* Quarteira	*S. aradensis* Arade	-0.016	0.031	-0.507	0.694
				-0.032	0.039	-0.827	0.796
				-0.023	0.026	-0.901	0.816
				0.029	0.025	1.138	0.128
				-0.054	0.033	-1.626	0.948
				-0.100	0.049	-2.014	0.978
				-0.051	0.053	-0.955	0.830
				0.020	0.027	0.762	0.223
				-0.032	0.038	-0.845	0.801
				-0.063	0.039	-1.642	0.950
	S. pyrenaicus Lizandro			-0.037	0.037	-1.000	0.841
				0.005	0.030	0.157	0.438
				0.057	0.033	1.723	0.042
				-0.029	0.039	-0.753	0.774
				0.030	0.032	0.939	0.174
				-0.044	0.044	-1.003	0.842
				-0.023	0.034	-0.680	0.752
				-0.053	0.048	-1.103	0.865
				0.002	0.045	0.049	0.480
				0.018	0.030	0.611	0.271
	S. pyrenaicus Canha			-0.009	0.035	-0.253	0.600
				-0.005	0.023	-0.219	0.587
				0.012	0.037	0.333	0.370
				-0.037	0.030	-1.232	0.891
				0.013	0.031	0.404	0.343
				-0.017	0.028	-0.612	0.730
				0.044	0.030	1.495	0.067
				0.050	0.017	2.907	0.002
				0.048	0.052	0.922	0.178
				0.130	0.015	8.524	0.000

Supplementary Table S12 (continued)

P1	P2	P3	P outgroup	D	sd	z	p-value
S. carolitertii	*S. pyrenaicus* Ocreza	*S. pyrenaicus* Almargem	*S. aradensis* Quarteira	0.043	0.032	1.325	0.093
				0.003	0.038	0.078	0.469
				0.020	0.025	0.798	0.212
				0.047	0.027	1.745	0.040
				-0.038	0.034	-1.138	0.873
				-0.079	0.047	-1.679	0.953
				-0.004	0.038	-0.094	0.537
				0.064	0.027	2.348	0.009
				0.016	0.038	0.432	0.333
				-0.001	0.036	-0.021	0.509
	S. pyrenaicus Lizandro			-0.014	0.041	-0.345	0.635
				0.039	0.028	1.410	0.079
				0.070	0.033	2.104	0.018
				-0.019	0.032	-0.600	0.726
				0.031	0.030	1.042	0.149
				0.006	0.038	0.152	0.440
				0.011	0.033	0.319	0.375
				0.016	0.043	0.372	0.355
				0.016	0.045	0.364	0.358
				0.013	0.025	0.534	0.297
	S. pyrenaicus Canha			0.031	0.030	1.001	0.158
				0.023	0.024	0.962	0.168
				0.018	0.038	0.482	0.315
				0.001	0.029	0.041	0.484
				0.042	0.031	1.342	0.090
				0.020	0.032	0.609	0.271
				0.046	0.030	1.557	0.060
				0.074	0.020	3.750	0.000
				0.062	0.043	1.444	0.074
				0.030	0.017	1.755	0.040

Supplementary Table S12 (continued)

P1	P2	P3	P outgroup	D	sd	z	p-value
S. carolitertii	*S. pyrenaicus* Ocreza	*S. pyrenaicus* Quarteira	*S. aradensis* Quarteira	-0.006	0.032	-0.191	0.576
				-0.040	0.035	-1.147	0.874
				-0.029	0.026	-1.118	0.868
				0.022	0.027	0.809	0.209
				-0.054	0.031	-1.721	0.957
				-0.088	0.046	-1.898	0.971
				-0.072	0.044	-1.629	0.948
				0.022	0.025	0.873	0.191
				-0.020	0.040	-0.505	0.693
				-0.057	0.034	-1.690	0.954
	S. pyrenaicus Lizandro			-0.040	0.038	-1.064	0.856
				-0.001	0.027	-0.038	0.515
				0.030	0.034	0.887	0.188
				-0.032	0.033	-0.985	0.838
				0.005	0.032	0.153	0.439
				-0.016	0.039	-0.417	0.662
				-0.008	0.031	-0.250	0.599
				-0.040	0.041	-0.976	0.835
				-0.001	0.043	-0.026	0.511
				0.022	0.030	0.748	0.227
	S. pyrenaicus Canha			-0.021	0.033	-0.634	0.737
				-0.007	0.020	-0.357	0.640
				0.015	0.036	0.421	0.337
				-0.014	0.026	-0.547	0.708
				0.004	0.028	0.135	0.446
				-0.009	0.027	-0.343	0.634
				0.035	0.030	1.154	0.124
				0.048	0.017	2.876	0.002
				0.042	0.052	0.820	0.206
				0.132	0.014	9.299	0.000

Supplementary Table S12 (continued)

Tree C

P1	P2	P3	P outgroup	D	sd	z	p-value
S. carolitertii	*S. pyrenaicus* Ocreza	S. pyrenaicus Lizandro	*S. torgalensis*	0.089	0.030	3.004	0.001
				0.103	0.029	3.597	0.000
				0.059	0.029	2.024	0.021
				0.116	0.024	4.828	0.000
				0.086	0.025	3.394	0.000
				-0.170	0.036	-4.782	1.000
				0.077	0.039	1.950	0.026
				0.098	0.024	4.048	0.000
				0.098	0.028	3.440	0.000
				0.058	0.037	1.552	0.060
		S. pyrenaicus Canha		0.105	0.027	3.913	0.000
				0.082	0.025	3.293	0.000
				0.079	0.025	3.211	0.001
				0.110	0.024	4.575	0.000
				0.088	0.023	3.914	0.000
				-0.143	0.033	-4.342	1.000
				0.063	0.035	1.820	0.034
				0.116	0.022	5.187	0.000
				0.092	0.028	3.266	0.001
				0.074	0.038	1.941	0.026

Supplementary Table S12 (continued)

P1	P2	P3	P outgroup	D	sd	z	p-value
S. carolitertii	*S. pyrenaicus* Ocreza	S. pyrenaicus Lizandro	*S. aradensis* Arade	0.080	0.032	2.517	0.006
				0.111	0.034	3.224	0.001
				0.066	0.034	1.958	0.025
				0.119	0.026	4.567	0.000
				0.081	0.031	2.646	0.004
				-0.181	0.044	-4.087	1.000
				0.076	0.047	1.618	0.053
				0.103	0.027	3.869	0.000
				0.116	0.038	3.078	0.001
				0.052	0.044	1.185	0.118
		S. pyrenaicus Canha		0.102	0.029	3.470	0.000
				0.082	0.032	2.563	0.005
				0.074	0.026	2.797	0.003
				0.109	0.026	4.181	0.000
				0.075	0.028	2.629	0.004
				-0.161	0.040	-3.989	1.000
				0.067	0.039	1.712	0.043
				0.119	0.025	4.735	0.000
				0.101	0.032	3.121	0.001
				0.077	0.043	1.773	0.038

P1	P2	P3	P outgroup	D	sd	z	p-value
S. carolitertii	*S. pyrenaicus* Ocreza	S. pyrenaicus Lizandro	*S. aradensis* Quarteira	0.078	0.032	2.456	0.007
				0.111	0.032	3.501	0.000
				0.058	0.031	1.852	0.032
				0.116	0.024	4.741	0.000
				0.062	0.028	2.212	0.013
				-0.184	0.040	-4.570	1.000
				0.055	0.040	1.384	0.083
				0.092	0.025	3.716	0.000
				0.111	0.035	3.172	0.001
				0.057	0.035	1.616	0.053
		S. pyrenaicus Canha		0.097	0.028	3.435	0.000
				0.086	0.027	3.239	0.001
				0.076	0.027	2.821	0.002
				0.107	0.025	4.372	0.000
				0.066	0.024	2.693	0.004
				-0.164	0.038	-4.362	1.000
				0.039	0.031	1.266	0.103
				0.111	0.022	5.096	0.000
				0.099	0.032	3.061	0.001
				0.081	0.036	2.246	0.012

Supplementary Table S12 (continued)

Tree D

P1	P2	P3	P outgroup	D	sd	z	p-value
S. pyrenaicus Lizandro	*S. pyrenaicus* Ocreza	*S. carolitertii*	*S. torgalensis*	0.009	0.024	0.371	0.355
				-0.025	0.025	-0.998	0.841
				-0.015	0.029	-0.512	0.696
				0.012	0.022	0.530	0.298
				-0.022	0.027	-0.804	0.789
				-0.167	0.039	-4.225	1.000
				-0.017	0.044	-0.386	0.650
				0.031	0.021	1.492	0.068
				-0.001	0.031	-0.045	0.518
				-0.019	0.035	-0.548	0.708
S. pyrenaicus Canha				0.008	0.025	0.343	0.366
				-0.010	0.025	-0.383	0.649
				-0.010	0.018	-0.584	0.720
				0.023	0.018	1.256	0.105
				-0.012	0.022	-0.564	0.713
				-0.182	0.037	-4.913	1.000
				-0.029	0.031	-0.939	0.826
				0.022	0.022	0.982	0.163
				-0.010	0.028	-0.359	0.640
				0.000	0.032	-0.002	0.501

Supplementary Table S12 (continued)

P1	P2	P3	P outgroup	D	sd	z	p-value
S. pyrenaicus Lizandro	*S. pyrenaicus* Ocreza	*S. carolitertii*	*S. aradensis* Arade	0.017	0.031	0.541	0.294
				-0.016	0.027	-0.587	0.721
				0.008	0.033	0.231	0.409
				0.023	0.023	0.995	0.160
				-0.013	0.033	-0.391	0.652
				-0.179	0.044	-4.028	1.000
				-0.030	0.053	-0.569	0.715
				0.027	0.026	1.012	0.156
				0.044	0.035	1.244	0.107
				-0.004	0.039	-0.105	0.542
S. pyrenaicus Canha				-0.009	0.030	-0.315	0.624
				-0.013	0.028	-0.486	0.687
				-0.014	0.019	-0.733	0.768
				0.021	0.019	1.053	0.146
				-0.023	0.027	-0.857	0.804
				-0.199	0.040	-4.937	1.000
				-0.051	0.038	-1.319	0.906
				0.024	0.023	1.064	0.144
				0.018	0.026	0.683	0.247
				-0.007	0.033	-0.212	0.584

P1	P2	P3	P outgroup	D	sd	z	p-value
S. pyrenaicus Lizandro	*S. pyrenaicus* Ocreza	*S. carolitertii*	*S. aradensis* Quarteira	0.010	0.031	0.308	0.379
				-0.016	0.023	-0.706	0.760
				0.003	0.032	0.080	0.468
				0.025	0.024	1.038	0.150
				-0.041	0.032	-1.290	0.901
				-0.175	0.040	-4.401	1.000
				-0.040	0.045	-0.891	0.814
				0.016	0.024	0.657	0.256
				0.042	0.035	1.205	0.114
				-0.004	0.035	-0.121	0.548
S. pyrenaicus Canha				-0.005	0.030	-0.163	0.565
				-0.010	0.023	-0.436	0.669
				-0.007	0.017	-0.387	0.651
				0.022	0.020	1.103	0.135
				-0.041	0.025	-1.616	0.947
				-0.193	0.036	-5.355	1.000
				-0.059	0.032	-1.842	0.967
				0.017	0.021	0.841	0.200
				0.019	0.027	0.684	0.247
				-0.002	0.031	-0.057	0.523